IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In the matter of:

Appl. No.	:	09/934,156 Confirmation No. 7387
Applicant	:	David Roth Rigney
Filed	:	August 21, 2001
TC/A.U.	:	1631
Examiner	:	Cheyne D. Ly
Response to	:	Office Action with mailing date of January 08, 2004

DECLARATION OF DAVID R. RIGNEY INCLUDING ATTACHMENTS 1, 2, AND 3

I, David R. Rigney, declare as follows:

1. I am the inventor named in the patent application referenced above. I make this Declaration based on my personal knowledge, and could and would testify competently to the facts stated herein.

2. My professional background has included graduate training in bioengineering, a doctorate in physics with a specialization in biophysics, postdoctoral training in biophysics, professional appointment with the title of biophysicist, appointment as Assistant Professor at Harvard University in the School of Medicine with a joint appointment at the Massachusetts Institute of Technology in the Division of Health Sciences and Technology, appointment on the staff of Boston Beth Israel Hospital, and Vice-President for Research and Development for the company GENETWORKS Inc.

3. While on the faculty of Harvard Medical School, I was the chairman of a faculty committee that was responsible for overseeing a resource that provided computer hardware and software support to molecular biologists who were on the faculty of Harvard Medical School and who had laboratories situated at Boston Beth Israel Hospital. In addition to acting in that

supervisory capacity, as a professional courtesy, I also personally provided hardware and software support to molecular biologist colleagues who worked in laboratories at Boston Beth Israel Hospital. I also ran my own cell and molecular biology laboratory in which I provided my own hardware and software support.

4. In connection with the duties described in paragraph 3 above, I joined a group called the Boston Area Molecular Biology Computer Types (BAMBCT). A description of BAMBCT is provided in **Attachment 1**, which is a copy of a web page "Welcome to BAMBCT" that I downloaded at http://genetics.mgh.harvard.edu/bambct/bambct-mission.html on June 18, 2004.

5. I believe that the members of BAMBCT (or its equivalent in areas outside of Boston) collectively constitute representative artisans for the art that is used in the patent application referenced above (self-described "molecular biology computer types" who provide hardware and software support to university molecular biology departments or BioTech companies).

6. Although I last had contact with BAMBCT in 1997 or 1998, I have no reason to believe that turnover of the members of the group is such that the range of backgrounds of the members of the group was any different at the time of the instant invention than it was in 1997 or 1998. If specification of a group of artisans is required for the period between 1997 or 1998 and the time of the instant invention, I would specify similarly self-described artisans in Austin, Texas over that time period, about whom I am personally familiar, but who apparently do not meet and confer as an organized group. The current description of BAMBCT shown in Attachment 1 is identical to my understanding of what BAMBCT was at all times that I was one of its members.

7. The BAMBCT mission statement shown in **Attachment 1** refers to a most frequent common denominator among members of BAMBCT, which is that most people run "GCG". This reference is to a software package known as the GCG Wisconsin Package. The GCG Wisconsin

package is described in **Attachment 2**, which is a copy of a web page "GCG Wisconsin Package" that I downloaded at the web site http://www.accelrys.com/products/gcg_wisconsin_package/ on June 19, 2004. The features described in **Attachment 2** are substantially the same as those described in 1998 in B.A. Butler, "Sequence Analysis Using GCG", Chapter 4 (pp. 74-97) in A.D. Baxevanis and B.F.F. Ouellette, eds., Bioinformatics. A Practical Guide to the Analysis of Genes and Proteins, New York: Wiley Interscience, 1998, except that the current user interface is more sophisticated. The 1998 chapter by Butler is provided in **Attachment 3** to this declaration.

8. Although I have never investigated in detail the backgrounds of the members of BAMBCT, my general impression is that about half of the group have doctorates in a technical subject and the other half have bachelors or masters degrees in a technical subject that requires some practical knowledge of computer hardware and software (computer science, physics, chemistry, engineering). The members of BAMBCT would ordinarily be able to write simple computer programs (scripts) for pre-existing software (like the GCG Wisconsin package) but would not in general be engaged in the writing of whole, compiled computer programs that require the design and implementation of a new algorithm. The members of BAMBCT would ordinarily be able to install off-the-shelf commercial software products, but would not necessarily be able to install non-commercial software without the debugging assistance of a non-biological computer systems analyst/programmer within their organization. The members of BAMBCT would ordinarily be experienced with sequence analysis as implemented, for example, in the GCG Wisconsin Package. BAMBCT members would not in general be technically experienced with the methods of gene expression analysis (Northern blots, microarrays, RT-PCR, etc.), although members would have some familiarity with the basics of such methods. I do not recall there ever being any mention, at a BAMBCT meeting or in BAMBCT email or later by any of the artisans in

Austin, Texas, about natural language processing, as described, for example, by Manning and Schutze (1999).

9. I believe that if "the artisan of ordinary skill in the art at the time of the instant invention" is taken to be a randomly selected member of BAMBCT or its equivalent in areas outside of Boston, then it would not have been obvious to that artisan to combine Andrade et al (1999) with McCallum (1998) to make the instant invention. This is primarily because that artisan would not be expected to have any prior training or experience in the technical aspects of McCallum (1998) such as the Naive Bayes concept; because that artisan would not necessarily be able to write a whole, compiled computer program that requires the design and implementation of a new algorithm; and because that artisan would not be able to conceive the convoluted sequence of changes needed to transform the combined Andrade et al and McCallum references, taken as a whole (including references cited by Andrade et al), into the instant invention, taken as a whole, without inadmissible hindsight.

10. The disclosure by McCallum states on its page 1, lines 6-7, that "Several of the examples also assume that you have downloaded the 20 newsgroups data set, unpacked them in your home directory, and therefore that its files are available in the directory ~/20_newsgroups." I performed this step as indicated above. I have also counted the words in each of the 20 groups, using the djgpp (unix) utility "wc" (word count). The number of words in the text corpus corresponding to each of the sample classes is as follows, as an indication of the size of the text corpus with which the program Rainbow is expected to work. I believe that this number of words is several orders of magnitude larger than the number of words to be found in the annotations or dictionary described in Andrade et al, so that there would not be a reasonable expectation of the useful or successful application of the Rainbow software described by McCallum to the annotations or dictionary of Andrade et al.

Class Name	Number of Words in the Text Corpus Correponding to that Class
alt.atheism	354053
comp.graphics	278585
comp.os.ms-windows.misc	234915
comp.sys.ibm.pc.hardware	216999
comp.sys.mac.hardware	203473
comp.windows.x	305914
misc.forsale	164281
rec.autos	237731
rec.motorcycles	217896
rec.sport.baseball	249160
rec.sport.hockey	301787
sci.crypt	348884
sci.electronics	225887
sci.med	313044
sci.space	310385
soc.religion.christian	404170
talk.politics.guns	356830
talk.politics.mideast	523816
talk.politics.misc	436764
talk.religion.misc	362082

The documents provided as Attachments 1-3 of this declaration are true and exact copies of what they are intended to represent. I declare under penalty of perjury under the laws of the United States of America that all the foregoing is true and correct.

Signed in Austin, Texas on July 7, 2004:

David R Rigney

David R. Rigney

CERTIFICATE OF EXPRESS MAIL UNDER 37 C.F.R. 1.10

I hereby certify that this paper is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service on the date indicated below and is addressed to Commissioner for Patents, P.O. Box 1450, Alexandria VA 22313-1450.

Printed Name: David R. Rigney

Date of Deposit: July 7, 2004

Signature: David R Rigney

Express Mail Label No. ER 826633934 US

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In the matter of:

Appl. No.	:	09/934,156	Confirmation No. 7387
Applicant	:	David Roth Rigney	
Filed	:	August 21, 2001	
TC/A.U.	:	1631	
Examiner	:	Cheyne D. Ly	
Response to	:	Office Action with mailing date of January 08, 2004	

ATTACHMENT 1 TO THE DECLARATION OF DAVID R. RIGNEY

Title of Attachment: Web page "Welcome to BAMBCT (Boston Area Molecular Biology Computer Types)"

Number of Pages: 1

Description of Attachment: Web page downloaded from the web site http://genetics.mgh.harvard.edu/bambct/bambct-mission.html on June 18, 2004

Welcome to BAMBCT (Boston Area Molecular Biology Computer Types).

We are both a real Bioinformatics mutual support group (meetings monthly at the best local mini-breweries for beer-drinking and discussions of hardware and software) as well as a virtual community of about 40 members communicating via email. On the average about 10 people attend monthly meetings at the local pubs. Most of us provide hardware and software support to university molecular biology depts or BioTech companies. The group started out with most people running GCG, but some now run DNA* and other programs or specialize in sub-areas of molecular biology computing. Some of us run DNA sequencing or synthesis facilities etc.

Access to the virtual community is via sending email to
bambct-list@molbio.mgh.harvard.edu
to directly get to all members on our mail exploder list.

You should post almost anything reasonable--job openings, requests for help, tips on new programs or Web sites, etc.

On occasion we have seminars or show-and-tell meetings to discuss issues many of us are interested in. We have had presentations on uses of Webservers in our Departments, large-scale DNA sequencing contig assembly, new GCG programs (presented by GCG), open source bioinformatics software.

Your suggestions are welcome.

Lance

Return to BAMBCT home page.

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In the matter of:

Appl. No.	:	09/934,156	Confirmation No. 7387
Applicant	:	David Roth Rigney	
Filed	:	August 21, 2001	
TC/A.U.	:	1631	
Examiner	:	Cheyne D. Ly	
Response to	:	Office Action with mailing date of January 08, 2004	

ATTACHMENT 2 TO THE DECLARATION OF DAVID R. RIGNEY

Title of Attachment: Web page "GCG Wisconsin Package"
Number of Pages: 15
Description of Attachment: Web page downloaded from
http://www.accelrys.com/products/gcg_wiconsin_package/ on June 19, 2004

accelrys

日本語サイト ●

ABOUT ACCELRYS SCIENCE INDUSTRIES [illegible] CUSTOMER DESK

WebStore | Site Map | Contact Us

Products

- Modeling/Simulation
- Informatics
- Client Services
- Consortia
- Desktop products
- Product finder
- Special offers
- WebStore

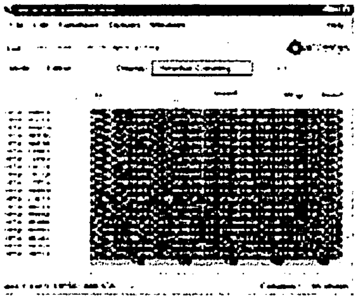
Home > **Products & Services** > **Informatics** > **GCG Wisconsin Package**

Search

GCG Wisconsin Package

On this page: Software Review / Interfaces / Advantages / Tour of Highlighted Programs / Requirements

Related Links: Version 10.3.1 Patch Now Available / What's New in 10.3 / DS SeqStore / Data Update Services / Transcription factor data files



SeqLab, free with the Wisconsin Package, provides a graphical interface to the Package's analysis tools plus project management capabilities. SeqLab's Editor (shown above) enables you to enter sequences, view multiple sequence alignments, as well as select the sequence ranges to analyze.

Molecular biologists worldwide use the GCG® Wisconsin Package® as their software of choice for comprehensive sequence analysis. The Wisconsin Package meets research needs across disciplines, project teams, and labs to provide an enterprise-wide solution. Based on published algorithms from the fields of mathematical and computational biology, the Package includes tools for:

- Comparison
- Database Searching and Retrieval
- DNA/RNA Secondary Structure
- Editing and Publication
- Evolution
- Fragment Assembly
- Gene Finding and Pattern Recognition
- Importing and Exporting
- Mapping
- Primer Selection

- Protein Analysis
- Translation

In addition to running within the UNIX operating system, the Wisconsin Package also runs on Intel-based x86 personal computers with Red Hat Linux 7.1 or 7.2. With the exception of the PAUP family of programs, the Linux edition of the Package provides the same functionality as the UNIX edition. PDF datasheet (A4, US Letter).

The Wisconsin Package is licensed on a per server basis. Contact your sales representative for more information.

A relational database version of the Wisconsin Package is available for use with DS SeqStore, our Oracle®-based data management and mining system. In addition to its sequence analysis capabilites, DS SeqStore includes tools to establish in-house relational databases; receive automated sequence data updates; and set up automated sequence analysis pipelines.

If your reseachers require access up-to-date public data behind the security of your institution's firewall, consider subscribing to one of our data update services. These services provide daily or bimonthly delivery of publicly-available sequence data already formatted for use with the Wisconsin Package.

Software Review

HMS Beagle—online magazine for the BioMedNet organization. Please follow this link (free registration required).

Interfaces

Three interfaces are available for the Wisconsin Package: SeqLab® and the command-line interfaces come with the Package while SeqWeb® is licensed separately.

SeqLab Supplied with the Package, SeqLab supplies a graphical user interface to the Wisconsin Package. SeqLab requires an X Windows display, such as an X server running on a PC or Macintosh, an X terminal, or a workstation that runs X Windows.

SeqLab provides an interactive sequence editor, convenient project management capabilities as well as a friendly interface for using Wisconsin Package programs. SeqLab supplies a rich visual display of sequences by individual bases or residues or by known sequence features that makes it easier to edit sequences or create and manipulate sequence alignments. In addition, you can click and drag to highlight multiple sequences or regions within sequences upon which to perform some analysis. SeqLab also makes it easy for you to annotate sequences due to your analysis or comparison with other sequences and their features. And if you have other programs that meet your needs, you can integrate them into SeqLab for ease of use and to create a common interface among them.

SeqLab's pull-down menus let you choose programs to manipulate the sequence(s) you have chosen. When you select a program from a pull-down menu, a separate window specific for that program appears. The program window includes a short message describing the program and presents all necessary input.

Command Line

The command-line interface enables you to run programs from the UNIX system prompt. All Wisconsin Package programs run in a similar manner; if you know how to run one, you know how to run them all.

Typical Scenario. Each program requires specific information to run successfully. When you run a program, it prompts you for an input file or asks you to answer questions with a yes or no or fill in a number or letter from a menu of available choices. The program suggests an answer for each prompt, except for the input file, allowing you to simply press <Enter> without typing a response.

Most programs have fewer than six prompts. Many program features are available as optional program modifiers. This design allows you to concentrate on the analysis that interests you without having to sift through many program modifiers each time you run the program.

SeqWeb

SeqWeb, an add-on product to the Wisconsin Package, allows you to connect to popular Wisconsin Package programs via Netscape® or Internet Explorer®. With SeqWeb, you can directly import files in a variety of formats; choose from a list of critical parameters for each program; link to databases on your local intranet or on the Internet from within program output; and run multi-program analyses in just one step. For more information, see SeqWeb.

Advantages

■ Well Established

Researchers worldwide use the Wisconsin Package and collaborate with Accelrys to provide programs that meet your needs.

■ Breadth of Analysis

The Wisconsin Package is the most comprehensive sequence analysis software available. Instead of using multiple software tools to achieve a final result, output from one Package program often acts as an input to others, providing a flow of analyses within a single interface.

■ Enterprise-Wide

Multi-user environment allows an unlimited number of scientists within your organization to share software and data.

■ Expertise

Our bioinformatics support staff is both highly rated and trained to provide scientific and technical expertise.

■ Current Data

We offer up-to-date access to the nucleic acid sequence databases GenBank and GenEMBL and the protein sequence databases PIR, SP-TrEMBL, SWISS-PROT, GenPept, NRL_3D, and Pfam.

■ Extendable Framework

Extensions enable you to plug other in-house or third-party software into SeqLab to provide a common interface and easy access.

■ Legacy of Commitment

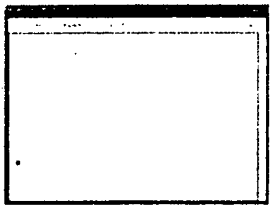
Since the beginning of bioinformatics, we have supported molecular biologists' research needs through software, education, and support, and continues to incorporate new technologies as this research discipline develops.

Tour of Highlighted Programs

The Wisconsin Package programs canvas a wide range of scientific interests, including sequence entry and fragment assembly, mapping, database searching, multiple sequence and evolutionary analysis, pairwise comparison, gene finding, DNA/RNA and protein secondary structure, translation, and display. Following are select Wisconsin Package programs grouped by function. Images are from the command-line interface. For a complete list of programs, click here.

Sequence Entry

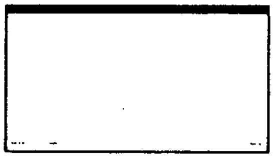
- SeqEd is an interactive editor for entering and modifying sequences.
- Individual sequences in Staden, EMBL, GenBank®, PIR®, IntelliGenetics, and FastA formats can be changed to Wisconsin Package format with the programs FromStaden, FromEMBL, FromGenBank, FromPIR, FromIG, and FromFastA. Sequences in other formats can be entered using the Reformat program.



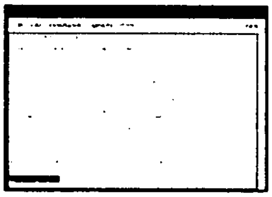
SeqEd enables you to view and change sequences.

Mapping

- Prime selects oligonucleotide primers for a template DNA sequence based on primer melting temperatures (Borer et al.), thermodynamic parameters for DNA (Breslauer et al.), PCR product melting temperatures (Baldino et al.), annealing temperatures of PCR primer pairs (Rychlik et al.), and self- or pair-annealing testing (Hillier and Green).



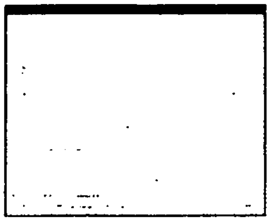
Prime selects oligonucleotide primers for a template DNA sequence.



Map displays enzyme

restriction sites as text.

- Map displays enzyme restriction sites above both strands of DNA along with protein translations below the DNA (Schroeder and Blattner).
- MapPlot displays restriction sites graphically.
- MapSort lists, by size, the fragments of single or multiple restriction enzyme digests.

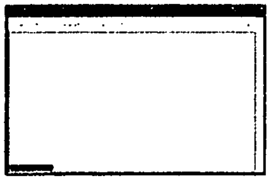


MapPlot displays enzyme restriction sites graphically.

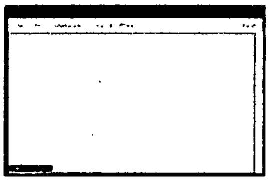
Fragment Assembly

A set of programs based on the methods of Staden let you enter, assemble, and view overlapping nucleotide fragments to create a single continuous sequence.

- GelMerge uses the method of Wilbur and Lipman to find overlapping regions among the fragments and the method of Needleman and Wunsch to align the fragments. A key option allows you to excise vector sequences.
- GelAssemble is a multiple sequence editor for viewing and editing contigs, or aligned assemblies of sequence fragments, assembled by GelMerge.
- GelView displays a schematic view of all contigs and their fragments in a project.
- GelDisassemble breaks up all contigs into their original fragments.



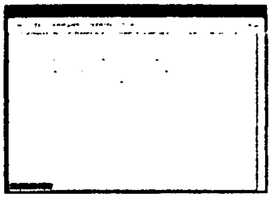
GelAssemble enables you to view and modify contigs.



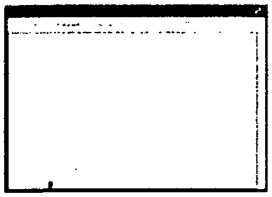
GelView displays contigs graphically.

Database Searching

- BLAST is a fast, statistically driven sequence searching and alignment tool (Altschul et al.) that can search databases on your computer or those maintained at the National Center for Biotechnology Information (NCBI) or at other regional BLAST servers.
- PSIBLAST uses position-specific scoring matrices (PSSMs) to score matches between query and database sequences, in contrast to BLAST, which uses pre-defined scoring matrices such as BLOSUM62. PSIBLAST may be more sensitive than BLAST, meaning that it may find distantly related sequences not found with a BLAST search
- FastA provides a more sensitive sequence searching and alignment tool (Pearson and Lipman). Variations of FastA include SSearch, TFastA, TFastX, and FastX.
- FindPatterns locates short ambiguous sequences like transcription factors in a database or set of sequences.
- Motifs searches sets of protein sequences or protein databases for the patterns defined in PROSITE (Bairoch).
- StringSearch searches through the sequence database references to locate sequences of interest, for example all human globin sequences.



BLAST searches for sequences similar to a query sequence.

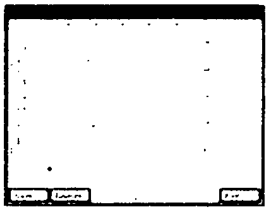


Motifs searches protein sequences for defined patterns.

- DataSet creates personal sequence databases for use by the database searching programs.

Pairwise Comparison

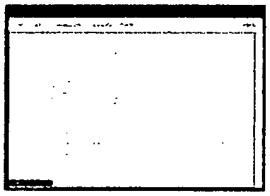
- DotPlot graphically displays an alignment between two sequences based on a number of matches within a given range (Maizel and Lenk) or complete matching over an entire short range (Wilbur and Lipman).
- BestFit finds the best segment of similarity between two sequences (Smith and Waterman).
- Gap finds the complete alignment of two sequences (Needleman and Wunsch).



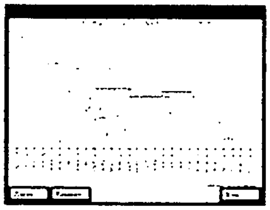
DotPlot graphically displays the alignment between two sequences.

Multiple Sequence Analysis

- PileUp creates a multiple sequence alignment of up to 500 sequences using the method of Feng and Doolittle, similar to the method of Higgins and Sharp. A dendrogram illustrating sequence similarity is also created using the strategy of Sneath and Sokal.
- ProfileMake creates a quantitative representation (a profile) of a family of aligned sequences that gives extra weight to parts of the alignment that are conserved across the family (Gribskov et al., 1987). The profile



PileUp creates a multiple sequence alignment for up to 500 sequences.



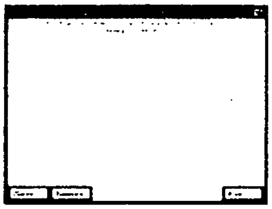
PileUp can also create a dendogram to graphically show sequence similarity.

can be used by ProfileSearch to search databases to find other members of the family (Gribskov et al., 1990).

- LineUp is an interactive editor for editing multiple sequence alignments.
- Pretty displays multiple sequence alignments.
- MEME (Multiple EM for Motif Elicitation - Timothy Bailey and Charles Elkan, University of California, San Diego) finds conserved motifs in a group of unaligned sequences and saves these motifs as a set of profiles. You can search a database of sequences with these profiles using the MotifSearch program.
- NoOverlap identifies the places where a group of nucleotide sequences do not share any common subsequences.

Evolutionary Analysis

- PAUPSearch provides a Wisconsin Package interface to the tree-searching options in PAUP (Phylogenetic Analysis Using Parsimony).
- PAUPDisplay provides a Wisconsin Package interface to tree manipulation, diagnosis, and display options in PAUP.
- Distances writes a matrix of the pairwise evolutionary distances between aligned sequences. To correct



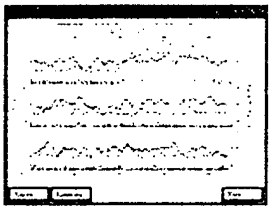
GrowTree creates a phylogenetic tree for a group of aligned sequences.

for multiple substitutions several methods may be chosen: for nucleic acid sequences, Kimura's two-parameter method (1980), the Tajima and Nei method, and the Jin and Nei method; for protein sequences, the Kimura method (1983); and for either type of sequence the Jukes and Cantor method.

- GrowTree creates a phylogenetic tree using neighbor-joining (Saitou and Nei) or UPGMA (Sneath and Sokal).

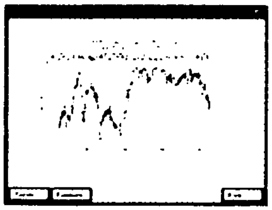
Gene Finding

- CodonPreference identifies and displays possible protein coding regions based on similarity of the codon usage in the sequence to a codon frequency table (Gribskov et al., 1984). Third position bias in the codon can also be displayed.



CodonPreference finds possible protein coding regions.

- TestCode uses the statistical method of Fickett based on the period three compositional constraints in the entire nucleic acid database to identify and display protein coding regions.



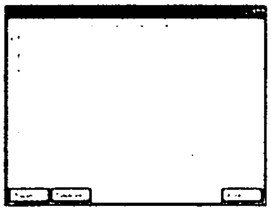
TestCode plots a measure of the non-randomness of the composition at every third base.

- Frames displays open reading frames for the six DNA translation frames.

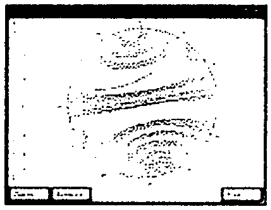
Frames displays open reading frames for the six DNA translation frames.

DNA/RNA Secondary Structure

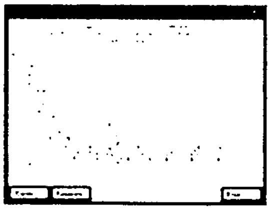
- MFold is an adaptation of the "mfold" package of Zuker and Jaeger which predicts optimal and suboptimal secondary structures for an RNA or DNA molecule (Zuker, and Jaeger et al.).
- PlotFold provides six ways to graphically display the optimal and suboptimal secondary structures calculated by MFold: energy dotplot, p-num, circles, dome, mountain, and squiggle) plots.



PlotFold's energy dotplot is a two-dimensional graph where both axes represent the same RNA sequence and each point on the graph indicates a base pair between the ribonucleotides whose positions in the sequence are the coordinates of that point on the graph.



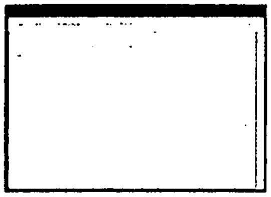
PlotFold's circles plot is a circular Nussinov graph of a nucleic secondary structure that shows the sequence as a segment of the circle.



PlotFold's squiggles plot represents the bonds formed between bases as chords.

Translation

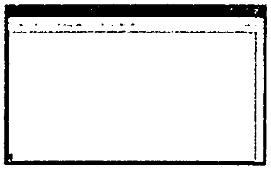
- Nucleotide sequences are translated into peptides using the Translate program.
- Peptide sequences are backtranslated into nucleotide sequences with the BackTranslate program.



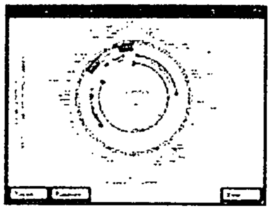
Translate translates nucleotide sequences into peptides.

Display

- Publish arranges sequence data for publication.
- PlasmidMap displays a circular plot of a plasmid construct.



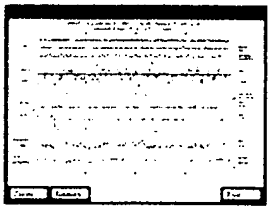
Publish produces publication-ready output.



PlasmidMap creates a circular map of a plasmid construct.

Protein Analysis

- PepPlot plots all of the standard measures of protein secondary structure: alpha-helix and beta-sheet prediction (Chou and Fasman, and Garnier et al.), hydrophobic moment

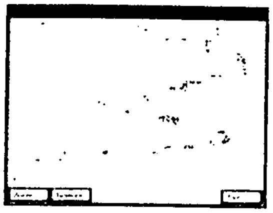


PepPlot plots all of the standard measures of

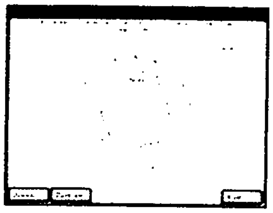
(Eisenberg et al.), and hydropathy (Kyte and Doolittle, and Engelman et al.).

- PlotStructure can display several standard measures of secondary structure as well as antigenic index (Wolf et al.) and surface probability (Emini et al.).
- HelicalWheel arranges the residues of a protein into a helix of adjustable angle and identifies hydrophobic residues.
- Isoelectric plots the charge of a peptide as a function of pH and calculates the isoelectric point.
- The Moment program helps find amphiphilic regions that coincide with an alpha-helix or beta-sheet structures (Eisenberg et al.)
- TransMem builds on the method of Sonnhammer et al. to predict likely transmembrane helices in one or more input proteins. The method is based upon a Hidden Markov Model (HMM) that has been trained on a set of membrane proteins with helical membrane spanning regions

protein secondary structure.



PlotStructure displays several standard measures of secondary structure, antigenic index, and surface probability.



HelicalWheel arranges the residues of a protein into a helix and identifies hydrophobic residues.

Requirements

Recommended Platforms for New Installations or Upgrades

Version 10.3 of the Wisconsin Package can be installed on a UNIX host system with users connecting to the host system via a modem or direct connection. The Package can also be installed on a personal computer running Red Hat Linux 7.1 or 7.2.

Computer	**Operating System**

Compaq	Tru64 UNIX 4.0E or later (Digital Unix) and 5.0
Silicon Graphics (RISC-Based)	IRIX 6.5
Sun (SPARC-Based)	Solaris 2.6, 7, or 8
IBM	AIX 5.1 and above
Intel x86 based Personal Computer	Red Hat/Linux 7.1 or 7.2

Memory and Storage

You need a minimum of 80 GB of disk space to maintain the Wisconsin Package and the complete set of databases as of March 2002. Since the databases have almost doubled in size in the past 12 months, plan on a comparable rate of increase as well as for possible significant increases due to the impending release of genomic data.

If you plan to update the data yourself (instead of purchasing one of our Data Update Services), you must convert the original distribution format of each supported database to GCG format for use by the Wisconsin Package programs. Therefore, double your storage requirements to accommodate the data in both formats.

On UNIX, system-wide recommendations include 128 megabytes (MB) core memory and 250 megabytes virtual memory. The Package must be installed using the C shell, although it can be run on both C shell and Korn shell for user programs. The operating system kernel must be configured to support the interprocess communication (IPC) mechanisms of shared memory and semaphores. The minimum shared memory size must be one megabyte or greater.

Note that as you increase the number of users and uses on your system, you also will need to increase memory.

Software CD Installation Set Sizes in MB

Set:	**Tru64 UNIX (Digital UNIX)**	**IRIX 64**	**IRIX32**	**Solaris**	**AIX**	**Linux**
Binary	43	49	44	76	48	35
Base	41	41	41	41	41	41
Total	84	90	85	117	89	76

Data Installation DVD Installation Set Sizes in MB (March 2002)

GenBank®	10,819	SWISS-PROT©	353
EMBL (Abridged)	217	SP-TrEMBL	1,012
NRL_3D Protein Structure	45	PIR®	461
GenBank Tags (EST and GSS)*	43,715	EMBL Tags (Abridged)*	36
BLAST	3,693	LookUp Indices	1,203
BLAST Tags*	4,422	Lookup Indices for Tags*	9,547
GenPept	617	Pfam	302

DataBasic Total: 18,722 MB
DataExtended Total: 76,442 MB

* Additional data supplied with the DataExtended service.

X Windows Software (Optional)

If you plan to use SeqLab or want to use X Windows as the graphics language to display Wisconsin Package graphics output, you need an X Windows server.

The User Workstation

The Wisconsin Package, mounted on a UNIX host system, probably is located remotely from the users who want to access it. You need either a modem or direct connection to the host system and need a terminal that preferably is able to display graphics. One or more locally available printers that can print plain text and graphics are also required. Plotters can be used for color graphic output if the terminal or printer does not have graphic capabilities.

We recommend using a graphics terminal. Approximately 20% of the programs provide output graphically. In addition, users may want to use SeqLab, the graphical user interface to the Package.

Terminals

A terminal can be provided in three different ways:

- *A microcomputer.* A PC or Macintosh can act as a standard terminal if terminal emulation software is installed. For the PC, we recommend SmarTerm, and for the Macintosh, VersaTerm Pro. Both provide Tektronix graphics for the screen as well as the ability to capture screen output for printing and editing. Files can be transferred between the host system and microcomputer. Note that other terminal emulators may or may not adequately function with the editors within the Package.

  A PC or Macintosh also can act as an X terminal when X server software is installed, allowing it to access SeqLab.
- *A UNIX workstation.* A window from the host system can be displayed and graphics as well as SeqLab would be accommodated.
- *A stand-alone terminal.* At minimum the terminal should be capable of sending and displaying standard text, for example the DEC VT series terminals. Ideally, the terminal should be able to display Tektronix (graphics terminal) or X Window (X terminal) graphics. X terminals, which can accommodate SeqLab, require that the host system have X Window Manager software installed.

Supported Graphics Terminals

X terminals
X Windows (for X servers)

Tektronix 4014
Tektronix 4105
Tektronix 4107
Tektronix 4207

VT330 (ReGIS)
VT340 (ReGIS)

Macintosh running VersaTerm Pro
PC running SmarTerm 340

Terminal Emulator Resource Addresses

SmarTerm
Esker, Inc.
465 Science Drive
Madison, Wisconsin 53711 USA
Tel: (608) 273-6000
Fax: (608) 273-8227
Web: http://www.esker.com

VersaTerm Pro
Synergy Software
2457 Perkiomen Avenue
Reading, Pennsylvania 19606 USA
Tel: (610) 779-0522
Fax: (610) 370-0548
Web: www.synergy.com/vt.htm

Microcomputer X Server Recommendations

See SeqLab requirements.

Printers and Plotters

Ideally, printers that can print text and graphics should be connected to the host system network and available in locations near users. Both PostScript and Hewlett Packard Graphic Language (HPGL) printers are supported for this purpose. The LPrint program in the Wisconsin Package prints ASCII files on PostScript laser printers.

Printers and plotters can be directly attached to terminals. Most programs write output files suitable for printing on any standard ASCII printer and many terminals have a pass-through port which can be used to connect a printer or plotter. If a terminal emulator will be used, extra setup may be needed to direct files on the host computer to a printer or plotter attached to the PC or Macintosh.

Color graphics can be displayed on color PostScript printers or on color HPGL printers and plotters.

To initialize a graphics device for printing or plotting, the Package provides the SetPlot program. A system manager defines the available local printing and plotting devices on the SetPlot menu, making it easy for users to access graphics terminals and printers.

Supported Printers and Plotters

Apple LaserWriter

DEC Printserver PS20
DEC LA210

HP LaserJet III
HP LaserJet IV
HP7475 plotter
HP7550 plotter

accelrys

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In the matter of:

Appl. No.	:	09/934,156	Confirmation No. 7387
Applicant	:	David Roth Rigney	
Filed	:	August 21, 2001	
TC/A.U.	:	1631	
Examiner	:	Cheyne D. Ly	
Response to	:	Office Action with mailing date of January 08, 2004	

ATTACHMENT 3 TO THE DECLARATION OF DAVID R. RIGNEY

Title of Attachment: Book Chapter Entitled "Sequence Analysis Using GCG"
Number of Pages: 24
Description of Attachment: B.A. Butler, "Sequence Analysis Using GCG", Chapter 4 (pp. 74-97) in A.D. Baxevanis and B.F.F. Ouellette, eds., Bioinformatics. A Practical Guide to the Analysis of Genes and Proteins, New York: Wiley Interscience, 1998

4

Sequence Analysis Using GCG

Barbara A. Butler

Genetics Computer Group, Inc.
Oxford Molecular Group
Madison, Wisconsin

INTRODUCTION

The advent of rapid, economical nucleic acid sequencing methods revolutionized many scientific disciplines including molecular biology, genetics, and biochemistry (Gilbert, 1981; Sanger, 1981). This technology also established a need for public databases to house the enormous amount of sequence information that was soon being generated in laboratories worldwide (Benson et al., 1997; Stoesser et al., 1997). The fields of bioinformatics and computational biology came of age with the establishment of these databases, since sequences submitted to them required analysis and annotation. In addition, existing database entries needed to be identified and retrieved by researchers wishing to study them further.

Bioinformatics can be described as the acquisition, analysis, and storage of biological information, specifically nucleic acid and protein sequences. Computational biology is the development of algorithms and computer programs integral to these endeavors. Both fields have grown dramatically in the past decade, driven by the enormous amount of data accumulating from whole-genome sequencing projects. Programs for analyzing sequences and searching databases are available from a number of sources, both commercial and academic. Packages for personal computers and Macintoshes are often expensive, especially for multiple users, and can lack a comprehensive array of programs for analysis and editing. Publicly available stand-alone (i.e., not part of a package) programs are inexpensive, in contrast to commercial programs, but they have to be downloaded and sometimes compiled on the local machine, and users have to become familiar with the format for input sequences and learn how to run them effectively. Network access to selected programs has become available recently, but it is difficult to perform analyses requiring more than one of these programs. For example, depending on the software used, a researcher can run a data-

Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins
Edited by A.D. Baxevanis and B.F.F. Ouellette
ISBN 0-471-19196-5, pages 74–97. Copyright © 1998 Wiley-Liss, Inc.

base sear
create an
This c
among a
access to
by the Ge
sin Packa
nucleic ac
full acces
an enviro
be expanc

Many
in other c
age and S
base entı
performe

THE W

The Wisc
sists of o
entries fr
lyzed wit
used as ir
mat that
modulari
ent ways
and desc
of all pro
The W
General i
ware requ
consin Pa

DATAB

GCG sup
with the
Wisconsi
searching
also prov
The tv
1997), pr
Database
These tw
hensive r

base search but cannot then align the sequences found by that search. It is also difficult to create an alignment of sequences and then edit that alignment.

This chapter introduces and discusses an environment that provides interoperability among a large number of sequence analysis and database searching programs as well as access to sequence data from a variety of sources. The environment is SeqLab®, developed by the Genetics Computer Group (GCG), as part of the Wisconsin Package™. The Wisconsin Package, a comprehensive set of sequence analysis programs, is distributed with public nucleic acid and protein databases. SeqLab is a graphical user interface (GUI) that permits full access to Wisconsin Package programs and supported databases. In addition, it provides an environment for creating, displaying, editing, and annotating sequences. SeqLab can also be expanded to include other publicly and locally available programs and databases.

Many of the analyses performed by Wisconsin Package programs are discussed in detail in other chapters of this volume, as are the databases distributed with the Wisconsin Package and SeqLab. Therefore, this chapter emphasizes the environment within which database entries and local sequences can be accessed, the types of analysis that can be performed, and the means of editing and annotating these entries and sequences.

THE WISCONSIN PACKAGE

The Wisconsin Package is a comprehensive sequence analysis software package that consists of over 120 individual programs, each performing a single analytical task. Database entries from public and private databases as well as individual sequence files can be analyzed with Wisconsin Package programs because there is a uniform format for sequences used as input to all programs. In addition, the output files from some programs are in a format that permits them to be further analyzed with other programs. Because of this, and the modularity of the package as a whole, a user can analyze sequences in a number of different ways by using programs in different combinations. The appendix of this chapter lists and describes the most widely used programs. A complete listing and detailed description of all programs can be found in the Program Manual for the Wisconsin Package.

The Wisconsin Package supports a number of UNIX platforms as well as OpenVMS. General information about GCG, the Wisconsin Package, supported platforms, and hardware requirements can be found on the GCG home page, */www.gcg.com/*, and in the Wisconsin Package User's Guide.

DATABASES THAT ACCOMPANY THE WISCONSIN PACKAGE

GCG supports and distributes five databases, two nucleic acid and three protein, for use with the Wisconsin Package. These databases are in both GCG format (for use with most Wisconsin Package programs), and BLAST format (for use with the BLAST database searching program). Indices for the LookUp program, for database reference searching, are also provided.

The two supported nucleic acid databases are the GenBank database (Benson et al., 1997), provided in its entirety, and an abridged version of the EMBL Nucleotide Sequence Database (Stoesser et al., 1997), consisting only of sequences not present in GenBank. These two databases have been combined for searching purposes into a single, comprehensive nucleotide database named GenEMBLPlus. This combined database includes the

GenBank and EMBL Nucleotide Sequence Database divisions for expressed sequence tag (EST), sequence tag site (STS), and genome sequence survey (GSS) entries. It is possible to search these three divisions separately with the specification TAGS or to search the GenEMBLPlus database without these divisions with the specification GenEMBL.

The three protein databases supported and distributed by GCG are the Protein Information Resource (PIR) International Protein Sequence Database (George et al., 1997), the SWISS-PROT Protein Sequence Databank (Bairoch and Apweiler, 1997), and SP-TrEMBL (Bairoch and Apweiler, 1997). SP-TrEMBL is a joint venture of the European Bioinformatics Institute and Dr. Amos Bairoch of the University of Geneva in Switzerland. It contains most of the predicted translated regions noted in EMBL database entries but does not contain any entries already present in SWISS-PROT. SP-TrEMBL entries are annotated using SWISS-PROT conventions, and as these entries appear in the SWISS-PROT database they will be removed from SP-TrEMBL. These two databases, SWISS-PROT and SP-TrEMBL, have been combined for searching purposes to create a comprehensive protein database named SWISS-PROTPlus.

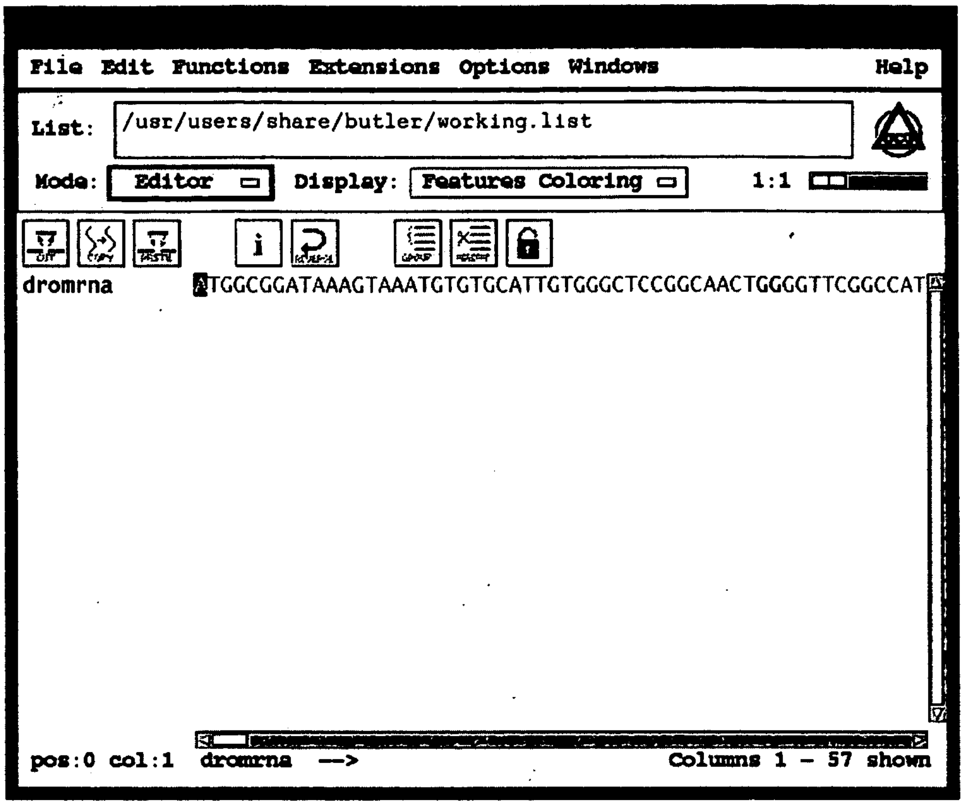
New releases of the databases supported by GCG are available bimonthly (following the GenBank database release schedule) as part of the GCG Database Update Service. Alternatively, Wisconsin Package utility programs and scripts are available for downloading and formatting database releases on site. These programs can also be used to update databases between releases or to format private data into databases for use with the Wisconsin Package. A list and description of these utility programs can be found in the Wisconsin Package System Support Manual. Databases in FASTA format can be used directly, without further formatting, with all programs included in the Wisconsin Package except the BLAST and LookUp programs.

THE SEQLAB ENVIRONMENT

SeqLab is a graphical user interface to the Wisconsin Package based on OSF/Motif™. It allows access to most Wisconsin Package programs and all supported databases from a Windows-based environment. Use of SeqLab requires an X-terminal or X-server software running on a microcomputer. Recommendations for X-server software can be found on the GCG home page, *www.gcg.com.*

After the Wisconsin Package has been initialized, SeqLab is launched from the UNIX prompt with the command `seqlab`. A window appears entitled SeqLab Main Window (Figure 4.1). There are two modes in which this main window can appear: Main List mode and Editor mode (referred to here as the "SeqLab Editor"). In Main List mode the SeqLab Main Window displays a list file containing the names of single-sequence, list, multiple-sequence format (MSF), and rich-sequence format (RSF) files as well as database entries. In Editor mode the SeqLab Main Window displays the sequences listed in these files and database entries. Users can toggle between the two modes with the Mode: option button on the SeqLab Main Window (Figure 4.1). Both modes permit access to Wisconsin Package programs and supported databases, but from the SeqLab Editor a user can also edit and annotate sequences. This chapter concentrates on the SeqLab Editor.

Across the top of the SeqLab Main Window is a menu bar; the menu options can be summarized as follows:



List: /usr/users/share/butler/working.list

Mode: Editor Display: Features Coloring 1:1

```
dromrna      ATGGCGGATAAAGTAAATGTGTGCATTGTGGGCTCCGGCAACTGGGGTTCGGCCAT
```

pos:0 col:1 dromrna ---> Columns 1 - 57 shown

Figure 4.1 The SeqLab Main Window in Editor mode.

File: Options for adding sequences from databases or directory files or for creating sequences de novo.

Edit: Options for moving and editing sequences and performing simple operations.

Functions: Wisconsin Package programs organized by analysis topics.

Extensions: A list of additional programs, if any, that can be run from within SeqLab.

Options: Preferences for displaying sequences and output, file management, and printing.

Windows: A list of windows for output display, program monitoring, and features annotation.

Help: Online help for Wisconsin Package programs and the SeqLab interface.

In addition to the Mode option button, the SeqLab Main Window includes a Display option button for changing the color or shading of sequences displayed and a scale bar for changing their horizontal scale. A panel of icons offers an alternative method for selecting editing options, viewing sequence information, and setting protections. The majority of the space in this window, however, is reserved for displaying sequences (Figure 4.1).

Adding Entries from Databases and Sequence Files from Directories

A sequence must appear in the SeqLab Main Window before it can be edited or analyzed with Wisconsin Package programs. Database entries are added either by entry name or by accession number. Single-sequence files (in GCG format), list, MSF, and RSF files are added by file name. (See the SeqLab Guide for details on these file formats and how they are created.)

To add an entry from a database to the SeqLab Main Window, use the left mouse button to select File from the menu banner and then Add Sequences From from the pulldown menu. Next, select Databases from the extended menu that appears. A Database Browser window will appear (Figure 4.2). Type the entry name or accession number of the desired database entry in the Database Specification text box at the bottom of the window. Click the Add to Main Window button and the Close button. This procedure can be abbreviated as follows. (Similar abbreviations are used throughout the chapter to describe keyboard and mouse commands.)

To add an entry from a database to the SeqLab Main Window:

1. Select File; go to Add Sequences From, and click Databases.
2. Type the entry name or accession number in the Database Specification text box of the Database Browser (Figure 4.2).
3. Click Add to Main Window, then Close.

Users can also add GCG-formatted sequence files to the list displayed in the SeqLab Main Window.

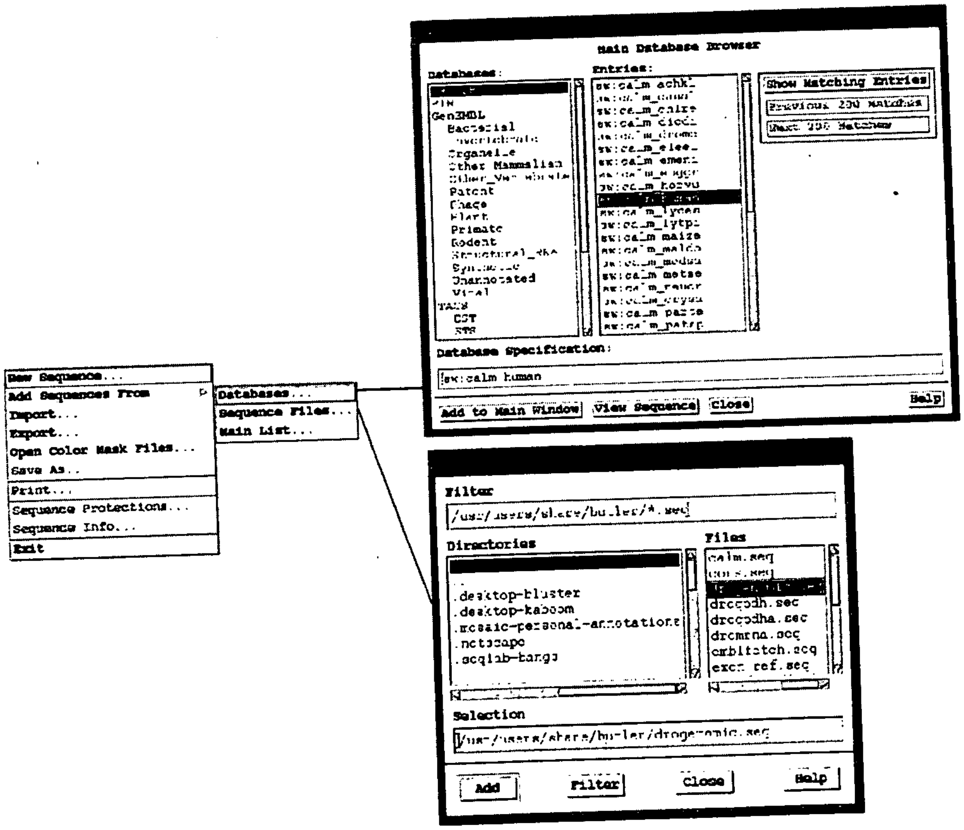


Figure 4.2 The Database Browser and Add Sequence windows for adding sequences to the SeqLab Editor.

To add

1. Sel
2. Sel
wil
the
3. Sel
4. Cli
5. Sel
6. Cli

Refere
double-c
Informat
necessar
ID/acces
Users
keys and
ing the n
gating w
are detai

Creatin

Users ca
To en

1. Se
2. C

When t
sequenc
ence inf
Sequen
the entr
erence i

Editin

It is im
played
when th
and pas
To c

1. S
2. S

To add a directory file to the SeqLab Main Window:

1. Select File; go to Add Sequences From, and click Sequence Files.
2. Select the appropriate filter in the Filter text box. (The default filter is *.seq, which will display all files in a directory ending in .seq. Replace *.seq with * to display all the files in a directory.)
3. Select the appropriate directory from the Directory area.
4. Click Filter.
5. Select the files by name from the Files area of the Add Sequences window.
6. Click Add, then Close.

Reference information for database entries and individual sequences can be viewed by double-clicking on the name of the entry or sequence. This action opens the Sequence Information window. Information in any of the text boxes on this window can be edited, as necessary. For example, it is often convenient to rename a database entry or add an ID/accession number to a sequence that is part of a large project.

Users can navigate within and among the sequences displayed in SeqLab with the arrow keys and horizontal and vertical scroll bars. Move to a residue within the sequence by typing the number of the residue and pressing the return key. Many other shortcuts for navigating within the SeqLab Editor, including moving relative to the current cursor position, are detailed in the SeqLab Guide.

Creating a New Sequence Entry

Users can also enter new protein or nucleic acid sequences into SeqLab.

To enter a new protein or nucleic acid sequence:

1. Select File and go to New Sequence.
2. Choose either DNA, RNA, or Protein from the New Sequence box.

When the listing appears, click at the beginning of the entry and either type in new sequence information or paste in sequence information from another window. Add reference information by double-clicking on the name of the new entry. This action opens the Sequence Information window. All the text boxes are editable, so in addition to renaming the entry, a description, author name, or ID/accession number can be included. General reference information can be added to the large text box at the bottom of the window.

Editing Existing Sequences

It is impossible to accidentally insert or delete residues because existing sequences displayed in the SeqLab Editor are protected. These protections can be changed, however, and when they have been removed, residues can be added and deleted, and it is possible to cut and paste sequences or regions of sequences between entries.

To change the protections on a sequence:

1. Select File and go to Sequence Protections.
2. Select all the buttons in the Sequence Protections window and click OK.

SeqLab is especially useful for editing multiple-sequence alignments. Useful features include the ability to move to absolute positions within either an individual sequence or an alignment, the ability to group and ungroup sequences so that a change in one sequence of the group also occurs in all other sequences of that group, and the ability to move "islands" of residues between gaps without changing the overall alignment. For example, a user can change an alignment that contains `gq...psqalt.......asw` to `gq......psqalt....asw` by sliding the `psqalt` island as if those six residues were beads on a string. The island overwrites one gap character to the right, as it moves in that direction, and a gap character appears to the left of the island such that the overall alignment is conserved. A complete list of editing operations is included in the Wisconsin Package SeqLab Guide.

ANALYZING SEQUENCES WITH OPERATIONS AND WISCONSIN PACKAGE PROGRAMS

Once sequences have been added and displayed in the SeqLab Main Window, they can be analyzed by running any Wisconsin Package program. The output files created by the programs are listed in the Output Manager window (see section entitled Viewing Output, below). Some of these files can be added back to the SeqLab Editor or SeqLab List mode for extended or related analysis. There are also a few simple operations that can be run directly from the SeqLab Editor.

Performing Simple Operations

The Edit menu in SeqLab Editor mode enables users to perform simple operations on displayed sequences without running programs. These operations include translating nucleic acid sequences, reversing and complementing nucleic acid sequences, calculating consensus sequences from aligned sequences, and finding short patterns. These operations have the advantage of running rapidly and displaying results automatically in the SeqLab Editor, where they can be edited, annotated, and, most importantly, used as input to Wisconsin Package programs selected from the Functions menu.

To select an operation:

1. Select a sequence by name or a range of a sequence.
2. Select Edit and go to the operation of choice.

Running Wisconsin Package Programs

Wisconsin Package programs are available for more extensive or robust analysis of sequences displayed in the SeqLab Editor. All the available programs are listed in the Functions menu and are divided by analysis topics. The Map program, from the Mapping functions topic, is used here as an example.

To run Map, a Wisconsin Package program:

1. Select a sequence by name or a region of a sequence with the cursor.
2. Select Functions and go to Mapping. Then select Map.

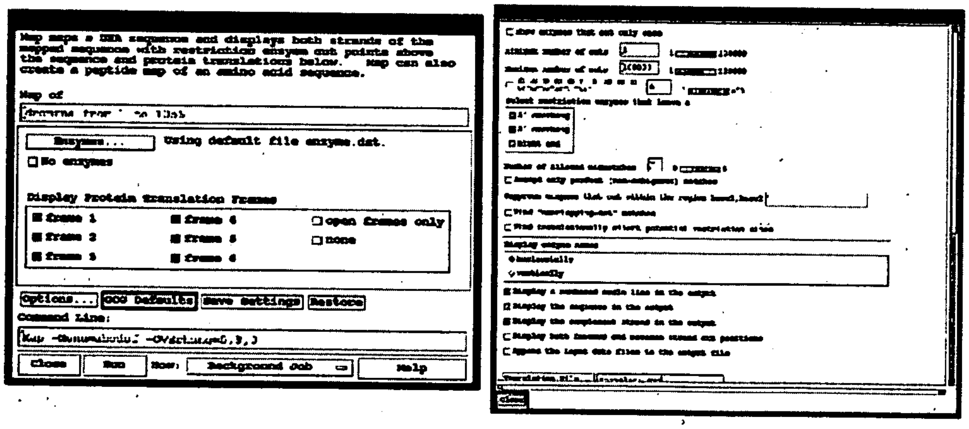
Selecting a program by name opens a Program window for that program. Every Program window has the same basic format, which includes the name of the selected sequence, the parameters required to run the program, a panel of buttons for selecting and saving optional parameters, and buttons for running the program, closing the window, and obtaining help. The Program window for the program Map is shown on the left in Figure 4.3.

Users can run a program with the default selections for required parameters or modify them with the buttons and text boxes on the Program window. In addition, each program has a unique set of optional parameters that will modify the analysis the program performs or change the way the output is displayed. These optional parameters are listed on the Program Options window, which is opened by selecting the Options button on the Program window. By selecting from required and optional parameters for the Map program, a user can select a subset of enzymes to include in a restriction map, opt for including only enzymes that produce a 5′ overhang on that map, or choose to omit the reverse complement strand normally included as part of a restriction map. The Map Options window is shown on the right in Figure 4.3.

Selecting the Run button on a Program window will run that program with the selected parameters and close the Program window. If a program is rerun during the same SeqLab session, the Program window will appear with all the previously selected parameters in place. Selected parameters can be saved between SeqLab sessions by selecting the Save Settings button. Selecting GCG Defaults from the Program window will reset the default parameter selections on both the Program and Program Options windows. All Program windows also include a Help button for accessing online help specific for that program.

VIEWING OUTPUT

Output files generated by programs run during a SeqLab session are listed in the Output Manager window (Figure 4.4).



Map maps a DNA sequence and displays both strands of the mapped sequence with restriction enzyme cut points above the sequence and protein translations below. Map can also create a peptide map of an amino acid sequence.

Map of

[illegible] from [illegible] to 1365

Enzymes... Using default file enzyme.dat.

□ No enzymes

Display Protein Translation Frames

■ frame 1 ■ frame 4 □ open frames only

■ frame 2 ■ frame 5 □ none

■ frame 3 ■ frame 6

Command Line:

Map [illegible]

☐ Show enzymes that cut only once

Minimum number of cuts [1] 1 10000

Maximum number of cuts [1(000)] 1 10000

[illegible]

Select restriction enzymes that leave a

■ 5′ overhang

■ 3′ overhang

■ blunt end

Number of allowed mismatches [0]

☐ Accept only perfect (non-ambiguous) matches

Suppress enzymes that cut within the region begin2,end2

☐ Find [illegible] matches

☐ Find translationally silent potential restriction sites

Display enzyme names

◆ horizontally

◇ vertically

■ Display a numbered scale line in the output

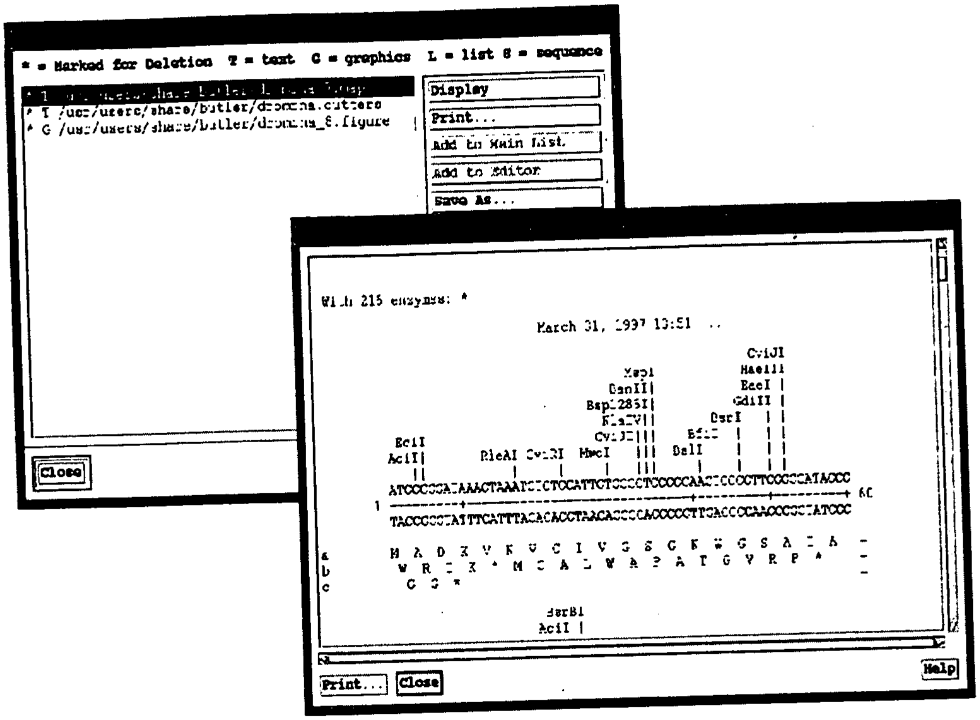
□ Display the sequence in the output

■ Display the complement strand in the output

☐ Display both forward and reverse strand cut positions

☐ Append the input data files to the output file

Figure 4.3 *Left:* Example of a Program window. For the Map program, this window is displayed by selecting Map from the Functions menu. *Right:* Example of a Program Options window. For the Map program, this window is displayed by selecting the Options button on the Map Program window.



```
* = Marked for Deletion  T = text  G = graphics  L = list  S = sequence
* T /usr/users/share/butler/dromina.cutters
* G /usr/users/share/butler/dromina_8.figure
```

```
With 215 enzymes: *

                    March 31, 1997 13:51  ..

                                                   CviJI
                              MspI               HaeIII
                            BsnII|               BceI |
                         Bsp1286I|              GdiII |
                            NlaIV||         BsrI   |  |
                           CviJI|||    BfaI  |     |  |
  BciI                                                      
  AciI|      BleAI CviRI  HhaI   |||  BslI   |     |  |
      ||          |     |     |  |||     |   |     |  |
   ATCCCCGATAAACTAAATCTCTCCATTCTCCCCTCCCCCAACTCCCCTTCCCCCATACCC
 1 ---------+---------+---------+---------+---------+---------+ 60
   TACCCCCTATTTCATTTACACACCTAACACCCCACCCCCTTGACCCCAACCCCCTATCCC

a  M  A  D  K  V  K  V  C  I  V  G  S  G  K  W  G  S  A  I  A  -
b   W  R  I  K  *  M  C  A  L  W  A  P  A  T  G  V  R  P  *   -
c    G  G  *                                                   -

                  BsrBI
                  AciI |
```

Figure 4.4 An output file created by the Map program displayed in an Output Display window (foreground). All output files created in SeqLab are displayed in the Output Manager window (background).

To open the Output Manager window:

1. Select Windows and go to Output Manager.

From this window output files can be displayed or printed. Click the Display button to display a highlighted file. An example of a displayed output file is shown in Figure 4.4. Click the Print button to send the selected file to a networked printer.

An output file generated in an earlier SeqLab session cannot be viewed or printed unless it is listed in the Output Manager window. Select the Add Text Files button, or Add Graphics Files, and select the file by name from the file browser that appears. Programs that produce graphics output create files with ".figure" extensions. When a file of this type is selected for display, it is translated for display in an X-window. When a file of this type is selected for printing, it is translated into either PostScript™ or HPGL™, depending on the printer selection and setup.

Some output files (sequence files, list files, and MSF files) can be added to the SeqLab Main List or Editor and used as input to Wisconsin Package programs. If such a file is selected in the Output Manager window, the Add to Main List and Add to Editor buttons will be active (Figure 4.4). If the selected output file cannot be added to these windows, the buttons will be inactive.

MONIT
TROUE

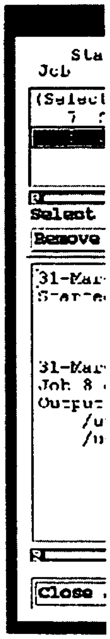
Every pro
ure 4.5).
Window.
To ope

1. Sel

The top h
ing the cu
the progr
this wind
is also po

ANNO
ANNO

A unique
ple, nucle



```
 Sta
Job
(Select
 7
Select
Remove
31-Mar
Starte
31-Mar
Job 8
Outpu
 /u
 /u
Close
```

Figure 4.5
window.

MONITORING PROGRAM PROGRESS AND TROUBLESHOOTING PROBLEMS

Every program run during a SeqLab session is recorded in the Job Manager window (Figure 4.5). This window can be accessed from the Windows menu bar on the SeqLab Main Window.

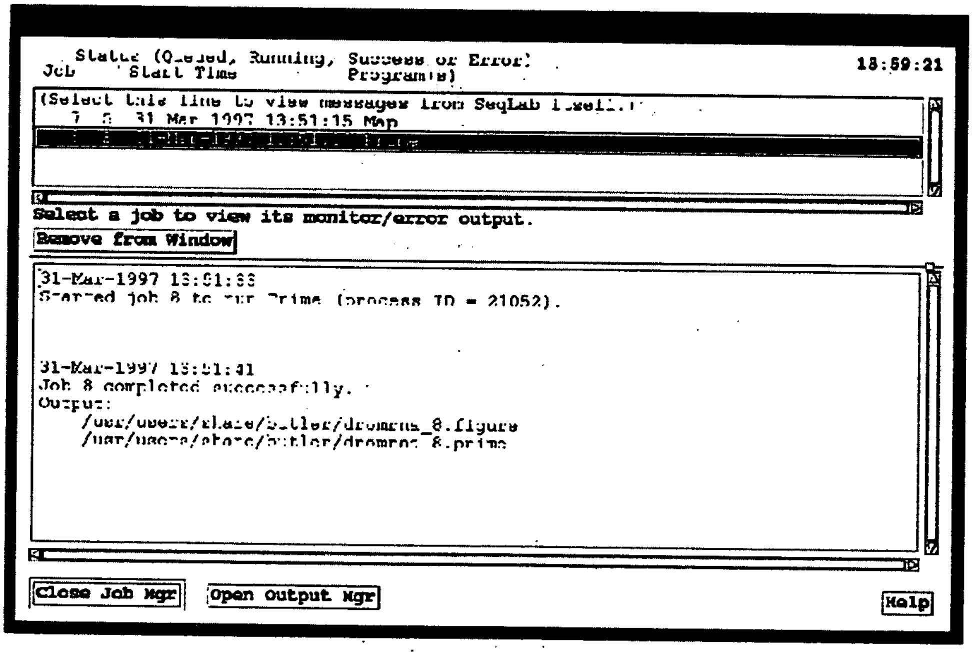
To open the Job Manager window:

1. Select Windows and go to Job Manager.

The top half of the Job Manager window is a log of all the programs that have been run during the current SeqLab session. The status of any program can be monitored by selecting the programs by name. If a program fails to run for any reason, a message will appear in this window and a log file for that program will appear in the Output Manager window. It is also possible to stop a running program from this window.

ANNOTATING SEQUENCES AND GRAPHICALLY DISPLAYING ANNOTATIONS IN THE SEQLAB EDITOR

A unique feature of SeqLab is its link to the Features table of database entries. For example, nucleic acid database entries often have features for the locations of coding regions,



```
    Status (Queued, Running, Success or Error)
Job    Start Time                Program(s)                    18:59:21
(Select this line to view messages from SeqLab itself.)
  7  S  31 Mar 1997 13:51:15 Map
[illegible]
Select a job to view its monitor/error output.
```

```
31-Mar-1997 13:51:33
Started job 8 to run Prime (process ID = 21052).



31-Mar-1997 13:51:41
Job 8 completed successfully.
Output:
    /usr/users/share/butler/dromrns_8.figure
    /usr/users/share/butler/dromrns_8.prime
```

Figure 4.5 The Job Manager window. All programs run during a SeqLab session are listed in this window.

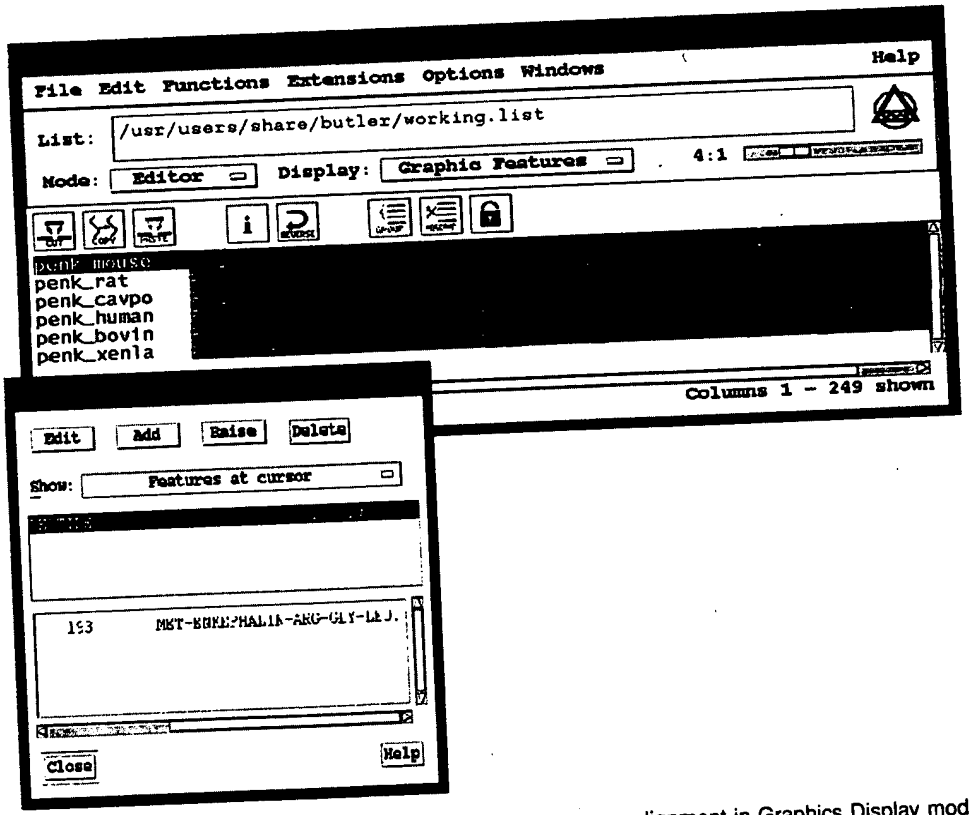
individual introns and exons, and polyadenlyation sites. SWISS-PROTPlus entries often have features for the locations of known protein pattern motifs, post-translational modification sites, and secondary structures. These features can be viewed in the SeqLab Editor with either colored residues (Features Coloring) or a schematic (Graphic Features).

To select features Display options:

1. Select the Display option button and then Features Coloring.
2. Select the Display option button and then Graphics Features.

An example of a Graphics Feature display for a set of aligned database entries is shown at the top of Figure 4.6. The 1:1 slide bar in the SeqLab Main Window (Figure 4.1) can be used to vary the horizontal scale of the schematic.

Database features can be displayed for an entry by selecting the Windows menu and then Features. This action opens a Sequence Features window (Figure 4.6). Users can opt to view all features for a sequence or just the selected feature. Selecting a feature in the upper area of the Sequence Features window displays detailed information about that feature in the lower area. This window can also be opened by double-clicking on a feature in an entry.



List: /usr/users/share/butler/working.list

Mode: Editor

Display: Graphic Features

4:1

penk_mouse
penk_rat
penk_cavpo
penk_human
penk_bovin
penk_xenla

Columns 1 - 249 shown

Show: Features at cursor

153 MET-ENKEPHALIN-ARG-GLY-LEU.

Figure 4.6 The SeqLab Editor displaying a multiple-sequence alignment in Graphics Display mode (*top*) and information about features are displayed in a Sequence Features window (*bottom*). Features can be added or edited by selecting option buttons at the top of this window.

Anoth
features :
tor wind
To ad

1. Hi
fo
2. Se
3. Se
4. Se
5. Ty
6. Ty
7. Cl

To ed

1. Se
2. Se
3. Se
4. M
5. Cl

SAVIN

When a
a rich-se
tures inf
informat
sequenc
ence anc
way are
stored ir

EXAM IN SE

Having :
tially to
been edi
is the at
mat the
solved v

Another unique and extremely useful feature of the SeqLab Editor is the ability to add features and edit existing ones. This is done from the Sequence Features and Feature Editor windows (Figure 4.6).

To add a feature:

1. Highlight a region with the cursor (or add ranges to the text boxes From and To, found in the Feature Editor).
2. Select Windows and then Features.
3. Select Add in the Sequence Features window.
4. Select Shape and Color in the Feature Editor window.
5. Type a name for the feature in the Keyword text box in the Feature Editor window.
6. Type a detailed comment in the Comments area of the Feature Editor window.
7. Click OK, then Close.

To edit a feature:

1. Select Windows and then Features.
2. Select the feature to edit in the Sequence Features window.
3. Select Edit in the Sequence Features window.
4. Modify Shape, Color, Range, Keyword, or Comments in the Feature Editor window.
5. Click OK, then Close.

SAVING SEQUENCES IN THE SEQLAB EDITOR

When a user exits SeqLab Editor mode, or saves editing work, the information is saved in a rich-sequence format (RSF) file. This is a new type of file that includes reference and features information as well as the sequence itself. The format of an RSF file enables features information to be displayed in the SeqLab Editor. RSF files can contain one or more sequence entries. If database entries are saved, copies of those entries (including all reference and features table information) are included in the RSF file. RSF files created in this way are automatically added to the current list file displayed in SeqLab List mode and are stored in the user's working directory.

EXAMPLES OF ANALYSES THAT CAN BE UNDERTAKEN IN SEQLAB

Having access to many sequence analysis programs confers the ability to use them sequentially to answer related questions or to repeat an analysis after the input sequences have been edited. The advantage of having access to both public databases and local sequences is the ability to use them both in a single analysis without first having to transfer or reformat them. This section describes six kinds of sequence analysis problems that can be solved with SeqLab.

Finding Open Reading Frames in Two mRNAs, Translating Them, and Aligning the RNAs and the Proteins

A user who has sequenced two related messenger RNAs may wish to find open reading frames, translate them, and create pairwise alignments of both the nucleic acid and amino acid sequences.

Add the sequences to the SeqLab Editor and run the Map program by selecting it from the Functions menu. The Map output file contains a restriction map and a display of open reading frames in any of the six possible translation frames. The begin and end positions of these open reading frames can be noted and selected as ranges in the sequences displayed in the SeqLab Editor where they can be translated with the Translate operation found in the Edit menu. These translations automatically appear in the SeqLab Editor.

Two related nucleic acid or protein sequences can be aligned to each other using either the Gap program (Needleman and Wunsch, 1970) or the BestFit (Smith and Waterman, 1981) program. Gap finds the best global alignment between the two sequences and is the program of choice if a user knows that the two sequences being compared are evolutionarily related. BestFit finds the best local alignment between two sequences and is the better program to use if the two sequences being compared are related not evolutionarily but, rather, functionally.

Finding Related Entries in Databases Through Reference Searching and Aligning Them

A user working with a member of a characterized sequence family may wish to find other members of that family and create a multiple sequence alignment of them all.

Select the LookUp program from the Functions menu. LookUp searches the reference sections of database entries for descriptive words and creates a list of the matching entries (Etzold and Argos, 1993; Etzold et al., 1996). Search for descriptive words in the Definition, Author, Keyword, and Organism fields of the reference sections and use the "and" (&), "or" (|), and "but not" (!) Boolean expressions between the words. For example, searching the Description field of SWISS-PROT entries for the words "lactate & dehydrogenase & h & chain" will create an output file listing lactate dehydrogenase H chain entries. This output file can be displayed from the Output Manager window and then added to the SeqLab Editor along with the user's sequence.

To create a multiple sequence alignment of all these sequences, select them by name and run the PileUp program from the Functions menu. The multiple-sequence file created by PileUp is also listed in the Output Manager window and can be added directly back to the SeqLab Editor. This step is recommended because Features table information from the database entries can be included with the alignment. The alignment can be edited if necessary, and similar features from the database entries can be added to the user's sequence, if present. The LookUp program window, the output file, and the alignment of the sequences within the output file are shown in Figure 4.7.

Searching a Database with a Query Sequence, Aligning the Found Entries and the Query Sequence, and Generating a Phylogenetic Tree

A user who has cloned and sequenced a gene of unknown function may wish to search a database for similar sequences. If any are found, the user may then wish to create a multiple-

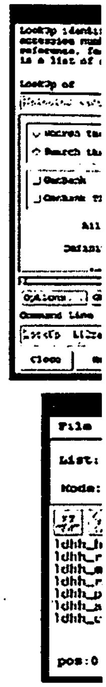
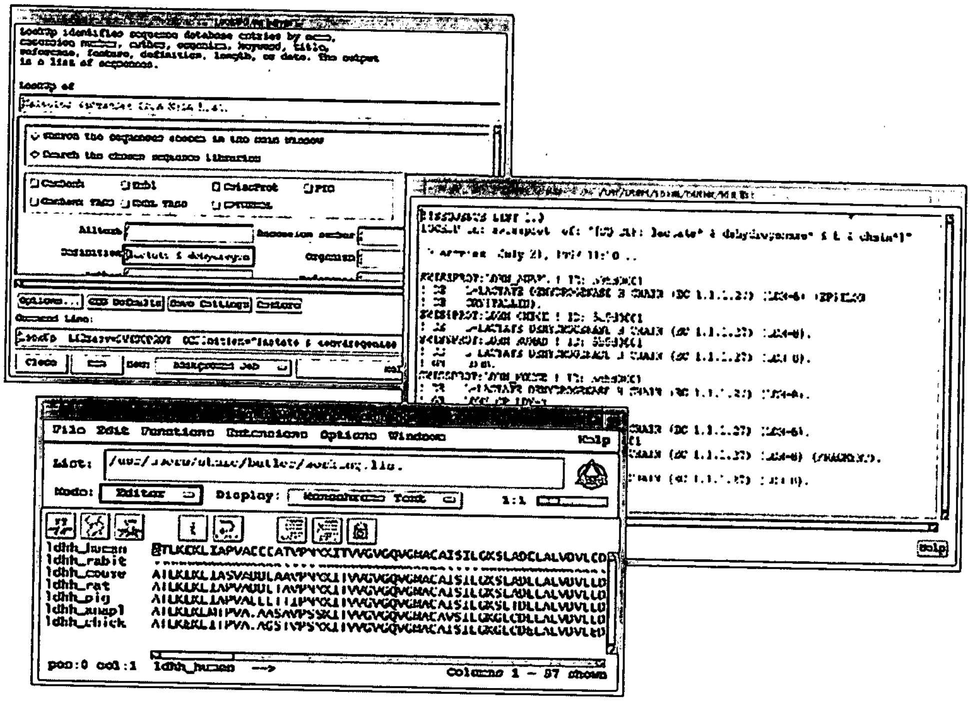


Figure 4.
from this s
window is
were adde
lower left-l

sequence
the data.
Add t
Function
similar t
window
ity betwe
only thos
play is d
Select
alignmen
dow and
alignmen
base entr
Select
GCG int
Parsimon
manipula



```
ldhh_human  BTLKDKLIAPVACCCATVPYVCITVVGVGQVGMACAISILGKSLADCLALVDVLCD
ldhh_rabit  ........................................................
ldhh_mouse  AILKDKLIASVADLLAAVPYVCLIVVGVGQVGMACAISILGKSLADLLALVDVLLD
ldhh_rat    AILKDKLIAPVADLIAVPYVCLIVVGVGQVGMACAISILGKSLADLLALVDVLLD
ldhh_pig    AILKDKLIAPVALLLIIAPYVCLIVVGVGQVGMACAISILGKSLIDLLALVDVLLD
ldhh_anap1  AILKDKLAIIPVA.AASAVPSSKLIVVGVGQVGMACAVSILGKGLCDLLALVDVLLD
ldhh_chick  AILKDKLIIPVA.AGSIVPSVCLIVVGVGQVGMACAISILGKGLCDLLALVDVLED

pos:0 col:1  ldhh_human  --->                        Columns 1 - 87 shown
```

Figure 4.7 Windows showing a database reference search using the LookUp program, the output file from this search, and a multiple-sequence alignment of the entries that were found. The upper left-hand window is the LookUp program window. The middle window displays the results of this search, which were added to the SeqLab Editor and aligned using the PileUp program. The alignment is shown in the lower left-hand window.

sequence alignment of the sequences most similar to the query, and generate a phylogram of the data.

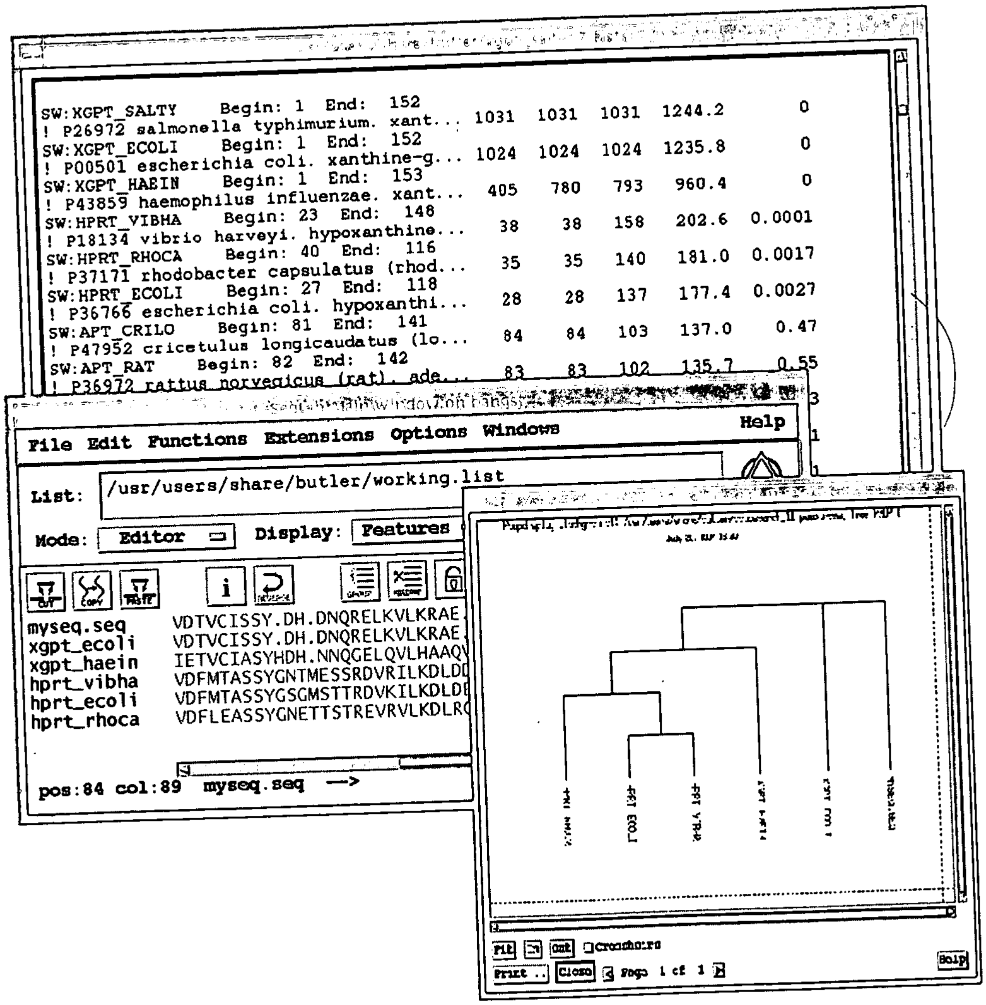
Add the query sequence to the SeqLab Editor and select the FASTA program from the Functions menu. FASTA (Pearson and Lipman, 1988) searches a database for sequences similar to a query sequence. The output file can be displayed from the Output Manager window and can be added directly to the SeqLab Editor. The best regions of local similarity between the database entries and the query sequence are noted in this output file, and only those regions of each database entry can be displayed in the SeqLab Editor, if a display is desired. Unwanted entries can be deleted from the SeqLab Editor altogether.

Select the PileUp program from the Functions menu to create a multiple-sequence alignment of these sequences. The output can be displayed from the Output Manager window and added to the SeqLab Editor, overwriting the existing, unaligned sequences. This alignment can be edited if necessary, and useful features table information from the database entries can be added to the query sequence.

Select the PaupSearch program from the Functions menu. This program provides a GCG interface with the tree-searching options in PAUP™ (Phylogenetic Analysis Using Parsimony) (Swofford, 1996). The PaupDisplay program provides a GCG interface to tree manipulation, diagnosis, and display options in PAUP. The output from the FASTA search, the alignment of the first six sequences, and the evolutionary tree generated from this alignment are shown in Figure 4.8.

Assembling Overlapping Sequence Fragments to Generate a Contiguous Sequence, Finding and Translating the Coding Regions of That Sequence, and Searching a Database for Similar Sequences

A user who has cloned a gene, subcloned it into a series of overlapping fragments, and sequenced those fragments may wish to reassemble the fragments into a contiguous



```
SW:XGPT_SALTY     Begin: 1  End:  152
! P26972 salmonella typhimurium. xant...  1031  1031  1031  1244.2       0
SW:XGPT_ECOLI     Begin: 1  End:  152
! P00501 escherichia coli. xanthine-g...  1024  1024  1024  1235.8       0
SW:XGPT_HAEIN     Begin: 1  End:  153
! P43859 haemophilus influenzae. xant...   405   780   793   960.4       0
SW:HPRT_VIBHA     Begin: 23  End:  148
! P18134 vibrio harveyi. hypoxanthine...    38    38   158   202.6  0.0001
SW:HPRT_RHOCA     Begin: 40  End:  116
! P37171 rhodobacter capsulatus (rhod...    35    35   140   181.0  0.0017
SW:HPRT_ECOLI     Begin: 27  End:  118
! P36766 escherichia coli. hypoxanthi...    28    28   137   177.4  0.0027
SW:APT_CRILO      Begin: 81  End:  141
! P47952 cricetulus longicaudatus (lo...    84    84   103   137.0    0.47
SW:APT_RAT     Begin: 82  End:  142
! P36972 rattus norvegicus (rat). ade...    83    83   102   135.7    0.55
```

List: /usr/users/share/butler/working.list

Mode: Editor Display: Features

```
myseq.seq     VDTVCISSY.DH.DNQRELKVLKRAE
xgpt_ecoli    VDTVCISSY.DH.DNQRELKVLKRAE
xgpt_haein    IETVCIASYHDH.NNQGELQVLHAAQ
hprt_vibha    VDFMTASSYGNTMESSRDVRILKDLD
hprt_ecoli    VDFMTASSYGSGMSTTRDVKILKDLD
hprt_rhoca    VDFLEASSYGNETTSTREVRVLKDLR
```

pos:84 col:89 myseq.seq —>

Figure 4.8 Windows showing a database search, an alignment of found database entries, and an evolutionary tree of this alignment. *Top:* Results of a FASTA search of the SWISS-PROT database. This file was added to the SeqLab Editor and the first six entries were aligned using the PileUp program. The alignment (*middle*) was used to create an evolutionary tree with the PaupSearch and PaupDisplay programs. The tree is shown in the lower window.

sequence. Once the contig has been assembled, the user may wish to find open reading frames within the sequence, translate them, and look for similar sequences in a database.

The programs of the Fragment Assembly System can be used to assemble overlapping sequence fragments. The GelStart program creates a project. The GelEnter program copies fragments into the project. The GelMerge program finds overlaps between the fragments and assembles them into contigs. The GelAssemble program is an editor for editing these contiguous units and resolving conflicts between the fragments. All these programs can be selected from the Functions menu. Once assembled, the consensus sequence for the final contig can be saved as a Sequence file and added to the SeqLab Editor.

Use the Map, Frames, TestCode (Fickett, 1982), or CodonPreference (Gribskov et al., 1983) programs to predict coding regions within the sequence. (All these programs can be selected from the Functions menu.) Use the Select Range function of the Edit menu to select the ranges predicted by these programs and the Translate operation of the Edit menu to translate them to protein. These proposed translated regions can also be added as features in the nucleic acid consensus sequence.

Select the protein sequence and then select BLAST (Altschul et al., 1990) from the Functions menu. BLAST searches databases for entries similar to a query sequence. Both remote and local searches are possible. The results can be displayed from the Output Manager window. If a local database is searched, the resulting file can be added to the SeqLab Editor or Main List window, allowing further analysis on the sequences found.

Aligning Related Protein Sequences, Calculating a Consensus Sequence for the Alignment, Identifying a Novel Pattern in the Sequences and Searching a Database for Sequences That Contain That Pattern, or Searching the Alignment Consensus for Known Protein Patterns

A user who has identified a group of related sequences may wish to align them and calculate a consensus sequence for the alignment. If a conserved pattern can be found in the alignment, the user may wish to search a database for other sequences that contain that pattern. The user may also wish to search the calculated consensus sequence for known protein patterns.

Select the sequences to align and select the PileUp program from the Functions menu to create a multiple sequence alignment. The PileUp output file can be displayed from the Output Manager window and added to the SeqLab Editor. It is possible for a user to realign a region of the alignment and place that region back into the original alignment. To do this, highlight the region and rerun PileUp. Select "realign a portion of an existing alignment" from the PileUp Options window. It might also be advantageous to select an alternate scoring matrix or different creation and extension penalties. The new output file will contain the original alignment, with the realigned region replacing the original alignment in that region.

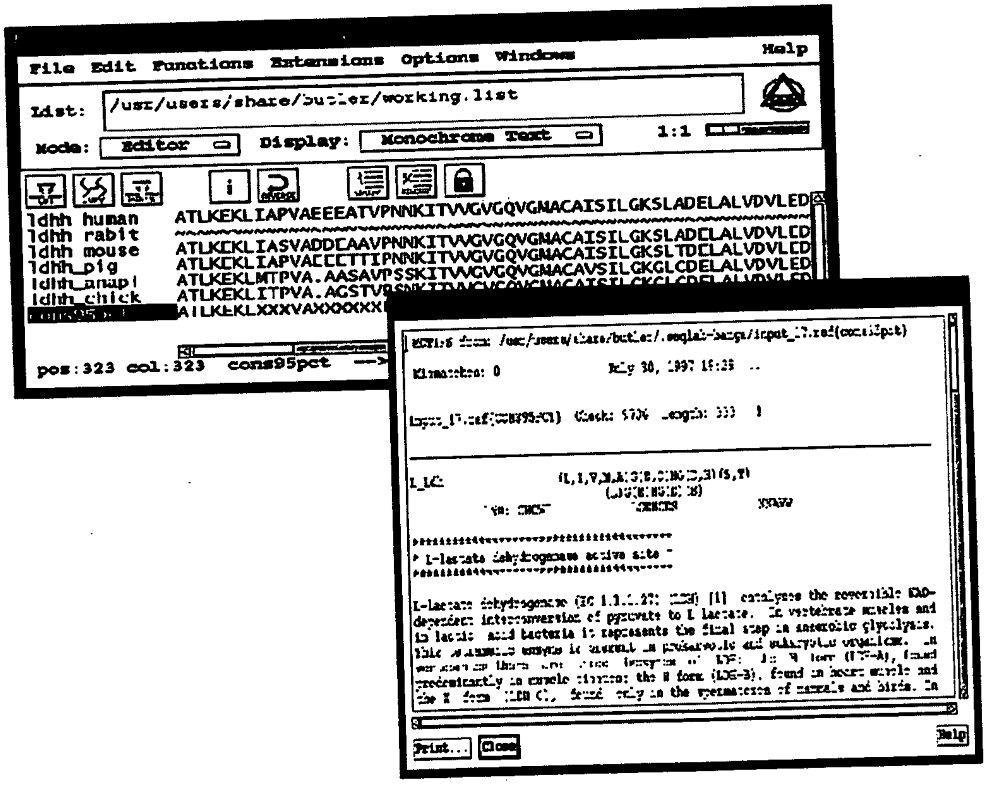
Calculate a consensus sequence for the alignment with the Consensus operation in the Edit menu. If a conserved pattern can be identified, select the FindPatterns program from the Functions menu. Cut the pattern from the consensus sequence, paste it into the FindPatterns Pattern Chooser, and search a database for sequences containing that pattern.

Alternatively, search the consensus sequence for known protein pattern motifs by running the Motifs program. Motifs searches protein sequences for the known protein patterns listed in PROSITE, the PROSITE Dictionary of Protein Sites and Patterns (Bairoch et al., 1997). If a motif is identified, add a feature to all the sequences, noting its position. An alignment of protein sequences plus a consensus sequence is shown in Figure 4.9, along with the results of a Motifs search.

Using Profiles for Similarity Searches and Aligning Related Sequences

A new and expanding region of sequence analysis is profile technology. A profile is a position-specific scoring matrix that contains information about all the residues at each position in a sequence alignment. This is in contrast to a consensus sequence, which contains only information about the consensus residue at each position. Once made, a profile can be used to search a database, database division, or search set for sequences similar to the sequences in the original alignment. It can also be used to align a single sequence to the alignment.

Use the ProfileMake program (Gribskov et al., 1987, 1990) to create a profile of a sequence alignment. Use the ProfileSearch program to search a database with the profile and the ProfileSegments program to display the results (Gribskov et al., 1987, 1990). Use the ProfileGap program to align a sequence to the profile (Gribskov et al., 1987, 1990). ProfileMake, ProfileSearch, ProfileSegments, and ProfileGap are all available from the Functions menu.



List: /usr/users/share/butler/working.list

Mode: Editor Display: Monochrome Text 1:1

```
ldhh human    ATLKEKLIAPVAEEEATVPNNKITVVGVGQVGMACAISILGKSLADELALVDVLED
ldhh rabit    ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
ldhh mouse    ATLKEKLIASVADDEAAVPNNKITVVGVGQVGMACAISILGKSLADELALVDVLED
ldhh_pig      ATLKEKLIAPVAEEETTIPNNKITVVGVGQVGMACAISILGKSLTDELALVDVLED
ldhh_anapl    ATLKEKLMTPVA.AASAVPSSKITVVGVGQVGMACAVSILGKGLCDELALVDVLED
ldhh_chick    ATLKEKLITPVA.AGSTV
cons95pct     AILKEKLXXXVAXXXXX

pos:323 col:323  cons95pct  -->
```

```
MOTIFS from: /usr/users/share/butler/.seqlab-[illegible]/input_17.rsf(cons95pct)

Mismatches: 0                  July 30, 1997 15:23 ..

input_17.rsf{CONS95PCT}  Check: 5706  Length: 333  !

________________________________________________________________

L_LDH             (L,I,V,M,A)(G,S,C)(N,D,S,H)(S,T)
                  [illegible]
        [illegible]            [illegible]            [illegible]

**********************************
* L-lactate dehydrogenase active site *
**********************************

L-lactate dehydrogenase (EC 1.1.1.27) (LDH) [1] catalyzes the reversible NAD-
dependent interconversion of pyruvate to L lactate. In vertebrate muscles and
in lactic acid bacteria it represents the final step in anaerobic glycolysis.
This [illegible] enzyme is [illegible] in prokaryotic and eukaryotic organisms. In
[illegible] there are three isozymes of LDH: the M form (LDH-A), found
predominantly in muscle tissues; the H form (LDH-B), found in heart muscle and
the X form (LDH-C), found only in the spermatozoa of mammals and birds. In
```

Figure 4.9 Windows showing a multiple-sequence alignment including a consensus sequence. The consensus was used to search PROSITE for known protein pattern motifs. An alignment of protein sequences in the SeqLab Editor, including a consensus sequence calculated at 95% identity, is shown in the upper window. This consensus sequence was used as input to the Motifs program to identify known protein pattern motifs common to the aligned sequences. The results from the Motifs program search are shown in the lower window.

EXTEN
NOT PA

Another k
ronment. I
gram to t
optional p
create a cc
It is not ne
Wisconsin
under the
SeqLab ar
ClustalW
9.0 of the
has been d
Progran
of the Seq

REFER

Altschul, S.
search to
Bairoch, A.
TrEMBI
Bairoch, A.
Acids Re
Benson, D.
1–6.
Chou, P. Y.,
amino ac
Etzold, T., a
put. App
Etzold, T., I
ogy data
Fickett, J. (
5303–53
George, D.
Sidman,
Barker, V
Sequence
Gilbert, W.
Gribskov, M
sis of pro
Gribskov, M
related p
Gribskov, M
146–159
Higgins, D.,
Alignme

EXTENDING SEQLAB BY INCLUDING PROGRAMS THAT ARE NOT PART OF THE WISCONSIN PACKAGE

Another key feature of SeqLab is the flexibility to insert additional programs in the environment. Briefly, the process entails obtaining an appropriate executable file for the program to be included and creating a configuration file that describes the required and optional parameters and formats the input and output files. Detailed instructions on how to create a configuration file can be found in the Wisconsin Package System Support Manual. It is not necessary to link these stand-alone program executables to any procedures in the Wisconsin Package. With this option, it is possible to run any program compiled to run under the operating system of the computer running the Wisconsin Package from within SeqLab and to view its output as easily as if it were part of the Wisconsin Package. ClustalW (Higgins et al., 1996) is the example extension program included with version 9.0 of the Wisconsin Package. Note that it is not a functional program unless the executable has been downloaded or built and the config file edited to point to the location of this file.

Programs added to the SeqLab environment can be selected from the Extensions menu of the SeqLab Main Window.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Bairoch, A., and Apweiler, R. (1997). The SWISS-PROT protein data bank and its supplement TrEMBL. Nucl. Acids Res. *25*, 31–36.

Bairoch, A., Bucher, P., and Hofmann, K. (1997). The PROSITE Database: Its status in 1997. Nucl. Acids Res. *25*, 217–221.

Benson, D. A., Boguski, M. S., Lipman, D. J., and Ostell, J. (1997). GenBank. Nucl. Acids Res. *25*, 1–6.

Chou, P. Y., and Fasman, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzymol. *47*, 45–147.

Etzold, T., and Argos, P. (1993). SRS—An indexing and retrieval tool for flat file data libraries. Comput. Appl. Bioscie. *9*, 49–57.

Etzold, T., Ulyanov, A., and Argos, P. (1996). SRS: Information retrieval system for molecular biology data banks. Methods Enzymol. *266*, 114–128.

Fickett, J. (1982). Recognition of protein coding regions in DNA sequences. Nucl. Acids Res. *10*, 5303–5318.

George, D. G., Dodson, R. J., Garavelli, J. S., Haft, D. H., Hunt, L. T., Marzec, C. R., Orcutt, B. C., Sidman, K. E., Srinivasarao, G. Y., Yeh, L. L., Arminski, L. M., Ledley, R. S., Tsugita, A., and Barker, W. C. (1997). The Protein Information Resource (PIR) and the PIR–International Protein Sequence Database. Nucl. Acids Res. *25*, 24–27.

Gilbert, W. (1981). DNA sequencing and gene structure. Science *214*, 1305–1312.

Gribskov, M., Devereux, J. D., and Burgess, R. R. (1983). The codon preference plot: Graphic analysis of protein coding sequences and prediction of gene expression. Nucl. Acids Res. *12*, 539–549.

Gribskov, M., McLachlan, M., and Eisenberg, D. (1987). Profile analysis: Detection of distantly related proteins. Proc. Natl. Acad. Sci. U.S.A. *84*, 4355–4358.

Gribskov, M., Luthy, R., and Eisenberg, D. (1990). Profile analysis. Methods Enzymol. *183*, 146–159.

Higgins, D., Thompson, J. D., and Gibson, T. C. (1996). Using CLUSTAL for Multiple Sequence Alignments. Methods Enzymol. *183*, 383–402.

Li, W. H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. *36*, 96–99.

Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol. *48*, 443–453.

Pamilo, P., and Bianchi, N. O. (1993). Evolution of the *Zfx* and *Zfy* genes: Rates and interdependence between the genes. Mol. Biol. Evol. *10*, 271–281.

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol. *183*, 63–98.

Pearson, W. R. (1996). Effective protein sequence comparison. Methods Enzymol. *266*, 227–258.

Pearson, W. R., and Lipman, D. J. (1988). Improved tools for biological sequence analysis. Proc. Natl. Acad. Sci. U.S.A. *85*, 2444–2448.

Sanger, F. (1981). Determination of nucleotide sequences in DNA. Science *214*, 1205–1210.

Smith, T. F., and Waterman, M. S. (1981). Comparison of bio-sequences. Adv. Appl. Math. *2*, 482–489.

Staden, R. (1980). A new computer method for the storage and manipulation of DNA gel reading data. Nucl. Acids Res. *8*, 3673–3694.

Stoesser, G., Sterk, P., Tuli, M. A., Stoehr, P. J., and Cameron, G. N. (1997). The EMBL Nucleotide Sequence Database. Nucl. Acids Res. *25*, 7–13.

Swofford, D. (1996). *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods)*, version 4.0 (Sunderland, MA: Sinauer Associates).

Zuker, M. (1989). On finding all suboptimal foldings on an RNA molecule. Science *244*, 48–52.

Wiscons
topics li
accessib
offers up

Pairwis

Gap: U
aligni

BestFit:
ment

FrameA
codor
neces

Compar
conta
result

ProfileM
quant
create

Multiple

PileUp:
sive, p
create

PlotSimi
multip

Databa

LookUp:
Numb
Date f

APPENDIX

Wisconsin Package programs are organized into topics based on scientific application. The topics listed are present in the SeqLab Functions menu. Most, but not all, of the programs accessible through SeqLab are listed, along with a brief description. The GCG home page offers up-to-date information and a complete list of Wisconsin Package programs.

Pairwise Comparison

Gap: Uses the algorithm of Needleman and Wunsch (1970) to find the optimal global alignment of two sequences.

BestFit: Uses the algorithm of Smith and Waterman (1981) to find the optimal local alignment of two sequences.

FrameAlign: Creates an optimal local alignment between a protein sequence and the codons in the three forward reading frames of a nucleotide sequence, adding gaps as necessary to maintain the reading frame.

Compare/DotPlot: Compares two protein or nucleic acid sequences, creates a file that contains information about the regions of similarity between them, and displays these results graphically as a dot matrix of similarity.

ProfileMake/ProfileGap: Creates a position-specific scoring table, called a profile, that quantitatively represents the information from a group of aligned sequences. ProfileGap creates an optimal alignment between a profile and a sequence (Gribskov et al., 1990).

Multiple Comparison

PileUp: Creates a multiple sequence alignment from a group of sequences using progressive, pairwise alignments. It also creates a graphic file showing the clustering used to create the alignment.

PlotSimilarity: Graphs the running average of the similarity scores of the sequences in a multiple sequence alignment.

Database Reference Searching

LookUp: Finds database entries by searching indexed fields such as Name, Accession Number, Author, Organism, Keyword, Title, Reference, Feature, Definition, Length, or Date for descriptive terms (Etzold and Argos, 1993).

Database Sequence Searching

BLAST: Searches a database for sequences similar to a query sequence (Altschul et al., 1990). The query and the database searched can be either peptide or nucleic acid in any combination. The program can search databases on an individual user's computer or databases maintained at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland.

FASTA: Searches a database for sequences similar to a query sequence. It was written by William Pearson and David Lipman (Pearson and Lipman, 1988).

TFASTA: Searches a nucleotide database for sequences similar to a protein query sequence. It translates the database sequences in all six frames before performing the comparison (Pearson and Lipman, 1988).

FrameSearch: Searches a nucleotide database or list file for sequences similar to a protein query. It can also search a protein database or list file for sequences similar to a nucleotide query. For each sequence comparison, the program finds an optimal alignment between the protein sequence and all possible codons on each strand of the nucleotide sequence, adding gaps to maintain the reading frame.

ProfileMake/ProfileSearch/ProfileSegments: ProfileMake creates a position-specific scoring table, called a profile, that quantitatively represents the information from a group of aligned sequences. ProfileSearch uses this profile to search a database, database division, or list file for sequences similar to those that created the profile. ProfileSegments displays the local regions of similarity between the database entries and the profile (Gribskov et al., 1990).

FindPatterns: Identifies sequences containing short patterns. Patterns can be defined ambiguously at each position and/or overall mismatching can take place.

Editing and Publication

Pretty: Varies the display of multiple-sequence alignments. It can also calculate a consensus sequence for the alignment.

Publish: Varies the display of single or multiple sequences. A menu of options for display, translating, and noting identities is provided.

MapSort/PlasmidMap: MapSort with the Plasmid option creates a file containing the locations of restriction enzyme recognition sites. This file can be graphically displayed with the PlasmidMap program Only circular restriction maps are possible.

Evolution

Distances/GrowTree: Creates a distance matrix of the pairwise corrected distances within a group of aligned sequences, expressed as a number of nucleotide or amino acid substitutions per 100 residues and constructs a phylogram.

PaupSearch: Provides a GCG interface to the tree-searching options in PAUP (Phylogenetic Analysis Using Parsimony) (Swofford, 1996).

PaupDisplay: Provides a GCG interface to tree manipulation, diagnosis, and display options in PAUP (Phylogenetic Analysis Using Parsimony) (Swofford, 1996).

Diverge: Estimates the number of synonymous and nonsynonymous substitutions per site between two nucleic acid sequences that code for proteins. It uses a variant of the method published by Li (Li, 1993; Pamilo and Bianchi, 1993).

Fragment Assembly

GelStart/GelEnter/GelMerge/GelAssemble: GelStart creates a fragment assembly project or initialized an existing one. GelEnter copies or enters fragments into the project. GelMerge finds overlaps between the fragments and assembles them into contigs, or contiguous regions. GelAssemble is an editor that displays the contigs for the resolution of conflicts between the fragments.

GelView: Displays all the contigs of a project at a given time and the names of all the fragments contained in each contig.

Pattern Recognition and Gene Prediction

TestCode: Uses algorithms developed by Fickett (1982) to predict protein-coding regions based on the nonrandomness of the composition of a nucleic acid sequence at every third base.

CodonPreference: Predicts protein coding regions based on codon usage and third position GC bias. Codon frequency tables for several organisms are available (Gribskov et al., 1983).

Frames: Graphically displays open reading frames for the six translation frames of a nucleic acid sequence based on the position of start and stop codons.

FindPatterns: Identifies sequences containing short patterns. Patterns can be defined ambiguously at each position and/or overall mismatching can take place.

Motifs: Finds known protein pattern motifs by searching protein sequences for the patterns defined in the PROSITE Dictionary of Protein Sites and Patterns (Bairoch et al., 1997).

Composition: Determines the composition of nucleic acid or protein sequence(s). For nucleotide sequence(s), it also determines dinucleotide and trinucleotide content.

CodonFrequency: Creates a codon frequency table from coding regions of sequences or existing codon usage tables. The output can be used with many Wisconsin Package programs including CodonPreference.

Importing/Exporting

Reformat: Formats sequence files, symbol comparison tables, or enzyme data files for use with Wisconsin Package programs. It can also be used to modify the display of sequences.

FromStaden: Converts a sequence file in Staden format (Staden, 1980) to GCG format. If multiple sequences are present in the file, individual sequence files will be created.

FromGenBank: Converts to GCG format a sequence file in GenBank flatfile format (Benson et al., 1997). If multiple sequences are present in the file, individual sequence files will be created.

FromPIR: Converts a sequence file in PIR format (George et al., 1997) to GCG format. If multiple sequences are present in the file, individual sequence files will be created.

FromFASTA: Converts a sequence file in FASTA format (Pearson and Lipman, 1988) to GCG format. If multiple sequences are present in the file, individual sequence files will be created.

ToPIR: Converts a GCG-formatted sequence file or files to PIR format (George et al., 1997).

ToFASTA: Converts a GCG-formatted sequence file or files to FASTA format (Pearson and Lipman, 1988).

ToStaden: Converts a GCG-formatted sequence file or files to Staden format (Staden, 1980).

Mapping

Map: Displays both strands of a nucleic acid sequence with restriction enzyme cut points above the sequence and protein translations below. Map can also create a peptide map of an amino acid sequence.

MapPlot: Graphically displays restriction enzyme recognition sites, one enzyme per line.

MapSort: Predicts the putative size of fragments after digestion of a nucleic acid with one or more restriction enzymes.

PeptideSort: Predicts the peptide fragments from digest of an amino acid sequence. It sorts the predicted peptides by weight, position, and relative retention times determined by high-performance liquid chromatography (HPLC). It also includes the composition of each peptide as well as a summary of the composition of the whole protein.

Primer Selection

Prime: Selects oligonucleotide primers for polymerase chain reaction (PCR) reactions, primer sequencing, and primer extension experiments. PCR is covered by U.S. Patents 4,683,195 and 4,683,202, owned by Hoffmann–LaRoche.

Protein Analysis

CoilScan: Locates coiled-coil segments in protein sequences.

HTHScan: Scans protein sequences for the presence of helix–turn–helix motifs, indicative of sequence-specific DNA-binding structures often associated with gene regulation.

Isoelectric: Predicts and plots a titration curve for a protein sequence.

ProfileScan: Uses a database of profiles to find motifs in protein query sequences (Gribskov et al., 1990).

PeptideSort: Predicts the peptide fragments from digest of an amino acid sequence. It sorts the predicted peptides by weight, position, and relative HPLC retention times. It also includes the composition of each peptide as well as a summary of the composition of the whole protein.

PepPlot: Predicts secondary structure using the method of Chou and Fasman (Chou and Fasman, 1978). The predictions are in a series of parallel plots. Plots for hydropathy and hydrophobic moment are included.

PeptideStructure/PlotStructure: Predicts and displays secondary structure antigenicity, flexibility, hydrophobicity, and surface probability for a protein sequence.

SPScai

RNA S

MFold
an R

StemL
min
bon

Transl

Transl

BackT
put
for

SPScan: Scans protein sequences for the presence of secretory signal peptides (SPs).

RNA Secondary Structure

MFold/PlotFold: Predicts and displays optimal and suboptimal secondary structures for an RNA molecule using the energy minimization method of Zuker (1989).

StemLoop: Finds stems, or inverted repeats, within a sequence. The user specifies the minimum stem length, minimum and maximum loop sizes, and the minimum number of bonds per stem.

Translation

Translate: Translates nucleotide sequences into peptide sequences.

BackTranslate: Translates an amino acid sequence into a nucleotide sequence. The output display helps the user to recognize minimally ambiguous regions that may be good for constructing synthetic probes.